DOCUMENT RESUME

ED 060 916 52 LI 003 607

AUTHOR	Maron, M. E.; Sherman, Don
TITLE	An Information Processing Laboratory for Education and Research in Library Science: Phase II. Final Report.
INSTITUTION	California Univ., Berkeley. Inst. of Library Research.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO	BR-7-1085
PUB DATE	Sep 71
GRANT	OEG-1-7-071085-4286
NOTE	121p.; (24 References)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Automation; Computers; Data Bases; Electronic Data Processing; *Information Processing; *Information Retrieval; *Library Education; *Library Science; Manuals; On Line Systems; Research
IDENTIFIERS	*University of California Berkeley

ABSTRACT

The results of the second 18 months (December 15, 1968-June 30, 1970) of effort toward developing an Information Processing Laboratory for research and education in library science is reported in six volumes. This volume contains: the introduction and overview, problems of library science, facility development and operational experience. One purpose of this volume is to clarify why an on-line laboratory for education and research in library science is desirable. The three major sections deal with: problems of education in library science, the organization of the laboratory, and the operation of the laboratory. (Other volumes of this report are available as LI 003608 through 003611). (Author/NH)

FINAL REPORT

Project No. 7-1085
Grant No. OEG-1-7-071085-4286

AN INFORMATION PROCESSING LABORATORY FOR EDUCATION AND RESEARCH IN LIBRARY SCIENCE: PHASE II

By
M.E. Maron
and
Don Sherman
Institute of Library Research
University of California
Berkeley, California 94720

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

September 1971

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

TABLE OF CONTENTS (cont.)

LIST OF FIGURES

This report contains the results of the second 18 months (December 15, 1968 - June 30, 1970) of effort toward developing an Information Processing Laboratory for research and education in library science. The work was supported by a grant (OEG-1-7-071085-4286) from the Bureau of Research of the Office of Education, U.S. Department of Health, Education, and Welfare and also by the University of California. The principal investigator was M.E. Maron, Professor of Librarianship.

This report is being issued as six separate volumes by the Institute of Library Research, University of California, Berkeley. They are:

- Maron, M.E. and Don Sherman, et al. An Information Processing Laboratory for Education and Research in Library Science: Phase 2.

  Contents--Introduction and Overview; Problems of Library Science; Facility Development; Operational Experience.

- Mignon, Edmond and Irene L. Travis. LABSEARCH: ILR Associative Search System Terminal Users' Manual.

  Contents--Basic Operating Instructions; Commands; Scoring Measures of Association; Subject Authority List.

- Meredith, Joseph C. Reference Search System (REFSEARCH) Users' Manual.

  Contents--Rationale and Description; Definitions; Index and Coding Key; Retrieval Procedures; Examples.

- Silver, Steven S. and Joseph C. Meredith. DISCUS Interactive System Users' Manual.

  Contents--Basic On-Line Interchange; DISCUS Operations; Programming in DISCUS; Concise DISCUS Specifications; System Author Mode; Exercises.

- Smith, Stephen F. and William Harrelson. TMS: A Terminal Monitor System for Information Processing.

  Contents--Part I: Users' Guide - A Guide to Writing Programs for TMS
  Part II: Internals Guide - A Program Logic Manual for the Terminal Monitor System

- Aiyer, Arjun K. The CIMARON System: Modular Programs for the Organization and Search of Large Files.

  Contents--Data Base Selection; Entering Search Requests; Search Results; Record Retrieval Controls; Data Base Generation.

Because of the joint support provided by the File Organization Project (OEG-1-7-071083-5068) for the development of DISCUS and of TMS, the volumes concerned with these programs are included as part of the final report for both projects. Also, the CIMARON System, whose development was supported by the File Organization Project, has been incorporated into the Laboratory operation and therefore, in order to provide a balanced view of the total facility obtained, that volume is included as part of this Laboratory project report. (See Shoffner, R.M., et al., The Organization and Search of Bibliographic Records in On-Line Computer Systems: Project Summary.)

ACKNOWLEDGMENTS

At one time or another since the inception of this project, over 150 different individuals (including students, faculty of the School of Librarianship, and staff members of the Institute of Library Research) were involved in and contributed to the design and actual building of the Information Processing Laboratory. Thus, although only a few of us are major authors of the volumes that make up this Final Report, we were assisted, in no small measure, by very many. And, it is my pleasure, as well as my responsibility, to acknowledge here with great thanks those others who collaborated on the work of this project.

From the very beginning and throughout most of the project, we had strong, constant support, advice, and encouragement from Dr. Ray Swank, who was Dean of the School of Librarianship.

Ralph Shoffner was Coordinator of Projects in the Institute of Library Research throughout the duration of this project, and his contributions were crucial. He was intimately involved in the design and development of the Laboratory. And his overall management and scheduling of the work was enormously effective. He was involved at all levels and in all facets of the work on the project ranging from long range planning to details of some of the software development. I gratefully acknowledge the great benefit that we obtained from his constant critical advice and energetic leadership. He labored long and hard on this project, and we owe him much.

Allan Humphrey was project manager for a good part of the existence of this project, and his constant, diligent attention to the work was essential. He made important creative contributions to the development of many of the computer programs, and he was excellent in inspiring and directing the work of still others.

Joseph Meredith was a key leader and contributor to a very large part of the Laboratory project. He was the driving force behind most of the work on the techniques for the on-line teaching of the subjects of reference and cataloging. A number of other important publications that resulted from Joe Meredith's work on this project testify to his creativity and productivity. This project owes much to his fine work.

Don Sherman, who co-authored this first volume of the Final Report, is another senior member of the Institute Staff who deserves special thanks for outstanding effort. For a large portion of the duration of this project, he was the manager of the operations of the Laboratory, and his efforts in administering, coordinating, synthesizing, and leading the work dealing with the use and evaluation of the Laboratory were heroic. He worked creatively and hard to organize and clarify much of the thinking about the operation of the Laboratory as an educational facility, and it is a genuine pleasure to acknowledge his fine contributions to this project.

A significant part of the Laboratory is its extensive inventory of computer programs, produced by various staff members of the Institute. The basic operating system, the Terminal Monitor System (TMS), is the work of Stephen Smith and William Harrelson. The Associative Search Program

(LABSEARCH) was designed and coded by C. Ravi. The MARC Search System (CIMARON) was the work of Arjun Aiyer. The Computer-Assisted Instruction Program, including the PILOT and DISCUS implementations, were programmed by Steven Silver and Rod Randall. REFSEARCH was coded by Allan Humphrey. These staff members worked hard; they were determined and dedicated. And without their efforts there would be no Laboratory.

Several doctoral students in the School of Librarianship made important contributions to this project. All of those who participated in the activities of this project were great; but the most intense, productive, intelligent, and sensitive work came from the heart, head, and hands of Ed Mignon. Another doctoral student who deserves special mention is Irene Travis, who not only co-authored with Mignon the volume on Labsearch, but also assisted in the using of the Laboratory to teach other graduate students the techniques of on-line association searching. Marcia Bates, Kelley Cartwright, Michael Cooper, Keith Stirling, and Ruth Patrick also contributed ideas, criticisms, and valuable insights.

An early contributor whose advice we acknowledge with great thanks is J.L. Kuhns.

There were several members of the faculty of the School of Librarianship whose advice and assistance were critical. In particular, it is a pleasure to acknowledge the early work of Mary Whouley, Grete Fruge, and Portia Swank. Another member of the faculty whose help we greatefully acknowledge is Victor Rosenberg.

At times we leaned very heavily on the members of the staff of the Computer Center. And we especially want to thank its Director Ken Hebert and two of his key associates Eric Sultan and William Vanderbeek.

And finally, it is a great pleasure to acknowledge with sincerest thanks the work of the office staff of the Institute. As part of the work on this project, we produced dozens of technical memos, technical papers, and reports, and those of the staff who helped produce all of this necessary documentation were simply great. Their intelligence, competence, and cheerfulness were a delight to encounter. Those deserving special thanks include: Carole Fender, Bettye Geer, Linda Herold, Jan Kumataka, Barbara Parrish, and Rhozalyn Perkins.

M.E.M.

1. INTRODUCTION

1.1 The Background

This is the Final Report, at the close of the second phase,* of a study initiated on July 1, 1967. The purpose of the study was to design, implement, and make an initial test and evaluation of an Information Processing Laboratory. The Laboratory was to serve students and faculty as a new type of facility designed specifically to enhance advanced education and research in the field of library science.

This project was funded primarily by the Bureau of Research of the Office of Education, but also by the University of California. The work over the past three years has been carried out by staff members of the Institute of Library Research (including graduate students representing over a dozen different disciplines ranging from electrical engineering and philosophy, to business administration and statistics), in collaboration with advanced students and some faculty members of the School of Librarianship, University of California at Berkeley.

1.2 Aims and First Questions

The central purpose of this project was to build a new kind of computer-based facility for advanced education and research in library science. No one before had built the kind of a Laboratory that we had in mind, and thus there were no blueprints that we could merely pick up and use. We lacked plans for how to proceed (because none existed), and furthermore, we ourselves lacked clarification of some very complex first questions such as:

- What ought to be the objectives of education in contemporary library science?
- What should be the relative emphasis between theoretical and applied library science?
- What does it mean to build a facility for education in library science?
- What are the major research objectives in library science?
- How would an on-line computer laboratory be relevant to education and research in library science?

We started with strong intuitive feelings of both what library science was about and what it ought to be about. We had strong feelings about the central role that the digital computer would play in the field of library science. We had ideas about the kinds of research and development activities that would have to take place successfully in order that the computer, in

*The Phase I Report was published in July 1969. See Maron, M.E., A.J. Humphrey, and J.C. Meredith, An Information Processing Laboratory for Education and Research in Library Science, Berkeley: Institute of Library Research, July 1969.

fact, be a significant force in library science. But successful research and development presupposes that there be a cadre of intelligent, properly ducated people to carry out that work. Thus, we were led to think about the kinds of education that would be relevant in this field. We started with questions, tentative assumptions, and some strong feelings about the field, the computer, and the future. Four years later, at this point in time, our ideas are clearer; our assumptions are better grounded; our understanding of library science is stronger.

At the same time that we were attempting to clarify basic questions and the nature of our long-range goals (and methods of achieving them), we found ourselves deep into the immediate problems of how to create an on-line laboratory. We were immersed in difficult problems of selection and use of hardware, design of complex software, planning and design of programs for teaching via display terminals, etc., and we began to realize how terribly complex and difficult it is to carry out this sort of a project. But we did persevere, and now there does exist a first version of an on-line laboratory for education and research in library science. In this Final Report we have attempted to set down not only the details of how this Laboratory is designed and operates, but also some thoughts relating to how we answer the fundamental questions that confronted us at the very beginning.

1.3 Current Status of the Laboratory

From the very beginning, our interpretation of the Laboratory has been that it be one that would allow a student (or staff member) to sit at a remote terminal and call up (from a central digital computer) data and analysis routines which would enable him to study, on-line, the properties of information search and retrieval techniques. Thus, in addition to its strictly physical aspects (e.g. terminals, modems, communications links, etc.), the Laboratory would consist of a variety of stored data bases upon which different search techniques could be exercised. And, of course, at the heart of the Laboratory, there would exist formal interrogation, search, and retrieval routines whose properties could be observed in use and thus studied by students via the remote terminals. A key idea from the very beginning has been that one can gain a new and deeper kind of understanding of formal information interrogation and search techniques by actually using these techniques on different corpora, and observing the consequences of their use in terms of what they retrieve. We have felt that the insights and understanding obtained by this sort of learning could not be duplicated via conventional lectures. And, we felt that this sort of a library science laboratory could provide a unique and valuable environment for empirical research on large files of bibliographic records. These ideas guided us and we built an on-line Information Processing Laboratory.

The physical equipment of the Information Processing Laboratory presently consists of three video display terminals connected by telephone line to a central digital computer. The remote terminals are controlled by a terminal monitor system which handles the communication between the terminals and the computer. There are four major "packages" that have been developed for use at these terminals. They are called: LABSRCH, REFSRCH, CIMARON, and

DISCUS. LABSRCH is an interrogation and retrieval system used for study of associative searching; i.e., the use of statistical techniques for measuring the closeness between index terms, and the use of these measures in automatic, on-line, literature searching. REFSRCH is a system for learning about the theory and practice of reference organization and search. CIMARON is a system for use in studying problems of interrogation and search of very large files of bibliographic records in L.C. MARC format. DISCUS is a language designed expressly for use in text oriented terminal-user interaction routines, such as computer-assisted instruction courses. In addition, we have under development a fifth set of programs which are designed for the on-line study of indexing.

The Laboratory has been in use on an experimental basis for several academic quarters, and we have had the preliminary experience of its use with over 150 students of the School of Librarianship.

1.4 Generalizations and Conclusions

It is difficult at this stage in this type of work to formulate a concise statement of major findings. Summary generalizations about the relationship between education (e.g., in library science) and complex technological systems (e.g., our Laboratory) tend to be either trivial or insufficiently backed up with supporting evidence. Specific observations and conclusions concerning LABSEARCH, REFSEARCH, DISCUS, and CIMARON are contained in each of the separate volumes of this Final Report, and they are summarized in the second part of this volume. However, at this point we want to make some remarks about the following question concerning the Laboratory: "Is it all worth it?"

It is very costly to develop such a Laboratory, and it is expensive to operate and use such a system. Our current estimate is that it costs approximately $20.00 per terminal hour to use the facility.

One of the major sources of complexity derives not merely from the system (e.g., the computer hardware, the interaction between system and user, or the structure of the search routines), but rather from its relationship to the organization of the material to be taught. The material in question deals with new concepts and techniques in library science. Our fundamental aim, of course, is education and the teaching of these concepts and techniques. In order to teach, the material must be unfolded, organized, and structured in certain logical patterns; and in our case, the material to be taught has to be structured to "fit" the "structure" of the Laboratory. We have only just begun to see how certain exercises should be organized and structured for best teaching via an on-line Laboratory. The problem is not due exclusively to the fact that we have a new educational facility, but due in large part also to the fact that we are dealing with material in library science that is complex, new, and changing. Again, the problem of how to teach certain topics is difficult under any circumstances, and the problem of how best to teach certain topics in library science via the Laboratory is still open.

We cannot say, in any definite way, how many terminal hours are needed to teach some of the concepts and techniques with which we are most concerned. Thus, we cannot say what the cost in dollars would be to teach via the Laboratory. But even if we could give this cost, it would not be good enough. Measuring the cost is less than half way toward answering questions about cost effectiveness. We don't know how to measure the benefit obtained by use of our Laboratory in educating advanced students in librarianship. (Can anyone measure the benefit of having a professor give a conventional lecture, or a demonstration on some given topic?)

Even assuming that we could give costs and a measure of benefit, how would one answer the larger question that we posed above; viz. "Is it all worth it?" The answer to this question depends on still other factors. That is, suppose we could show that by using our Laboratory we could improve education in librarianship sixfold, with a three-fold increase in cost. One must then ask: Who is it who benefits in the sixfold way? And, who is it who pays the three-fold cost? These days, we feel especially sensitive to the relevance of these kinds of questions, and there are continuing pressures to produce the answers. Our Laboratory is expensive to operate. We feel, but cannot prove, that the use of the Laboratory is, and can be increasingly, very beneficial in education and research in library science. We are still learning how best to use such a facility. We are not at all able to answer the question "Is it worth it?" But also, we should realize that we cannot provide quantitative data to answer this type of question for most of the things that we do.

1.5 Organization of the Final Report

The purpose of this Final Report is to describe, analyze, criticize, evaluate, and generally make explicit our experience in building and using the Information Processing Laboratory. The emphasis is not on the chronology of events leading from our first plans in 1967 to the evaluation of our current system. Rather, the emphasis is toward presenting a detailed picture of the Laboratory as it now exists and is being used for education and research. This series of volumes that make up the Final Report is the work of many hands; therefore, readers will find descriptions of our activities written at different levels of depth and detail. We have attempted to make this report very complete, yet easily accessible to those who may not want all the details. We hope that this report will be a useful document, not only as a guide to educators in library sciences elsewhere who might be involved in educational planning, but also to students and researchers who might want to use our Laboratory.

This Final Report is organized into six separate volumes including this one (Volume I) which acts as the overall introduction. There are, as we have said, four distinct major "packages" that presently constitute the logical inventory of the Laboratory. Each of these is aimed at providing a unique educational and research tool in library science. We have prepared users' manuals for each of these packages: LABSEARCH, REFSEARCH, DISCUS, and CIMARON. These are presented in four of the six volumes of this Final Report, (Volumes II, III, IV, and VI, respectively). A detailed description

of the monitor system (TMS Users' Manual) is contained as a completely separate volume (Volume V).

Volume I of this Final Report consists of three major sections (excluding this introductory section). Section 2, written by M.E. Maron, deals with problems of education in library science. Sections 3 and 4, written by Don Sherman, deal with the organization and operation of the Laboratory. One purpose of Volume I is to clarify and attempt to answer some first questions about why one would want an on-line Laboratory for education and research in library science. Thus, it seemed appropriate, before describing the Laboratory and our experience in using it, to set the stage for that discussion by attending to some questions about education and research needs in this field - with special emphasis on those aspects of education and research in library science that relate to the computer. If education, properly conceived of, is preparation for the future, then one can only describe what a relevant education should consist of by relating it to conjectures about the future of the field. Thus we are led to ask what role the computer will play in this field in the future. There will be many uses (and perhaps some misuses) for the computer in the service of library science. These uses range from the mechanization of some strictly clerical processes to automation of some aspects of information indexing and search, and perhaps one of the more important roles for the computer will be that of helping us to learn about how to teach new techniques for information searching and retrieval.

As a prelude to the detailed descriptions of the organization and operation of the Laboratory, we have attempted (in Section 2 of this volume) to unfold and develop the following motivating argument: the central problem of library science is the problem of access; viz., the problem of how to analyze and search in order to find needed information. How should searching be done? How can searching be mechanized? What are the theoretical principles and practical techniques for information searching? There exist practical techniques (both logical and technological), but there are few theoretical principles. We argue that because of the depth and complexity of the problem of information access, it may be decades (if ever) before any full theory is developed in operational terms (so that it can be implemented by a computer). However, improved search techniques can be uncovered, extended, refined, and taught in a problem solving laboratory -- in an on-line laboratory where techniques of access are used and where the consequences of those uses are made immediately available for analysis. The problem of information search is a complex type of problem solving activity, and not unlike certain other complex problem solving activities, it should be approached by developing proper searching tactics and strategies. Furthermore, the learning of tactics and strategies cannot be achieved by being told--but by doing. Learning skills of information searching must come from the systematic analysis of the activity of searching. A major objective of the Information Processing Laboratory is that it serve as an environment for the systematic study and learning of information search tactics -- of effective search techniques for use by people and machines.

2. PROBLEMS OF LIBRARY SCIENCE

2.1 Introduction

What is happening in contemporary library science? How rapidly is this field changing and developing? In what directions is it moving? When if ever, for example, will we have fully automatic libraries? What are the prospects for using computers to store millions of items of text, and then to analyze the stored text automatically in search of information that would be relevant to an inquiry? What can be said, realistically, about what the organization and operation of library systems will be like in another decade? Will we have any fundamental theories describing optimal principles for storing and retrieving information? What kinds of practical and theoretical education should today's library schools be offering to those students who will be the library scientists of the future? All of these very tough questions (and so many more) erupt immediately as soon as one thinks about designing "an information processing laboratory for education and research in library science." Where is one to start in the search for clarification of these (and related) issues? First, a word about terminology.

The term "library science" is awkward to some; to others the expression is pretentious and misleading. The term is pretentious and misleading if "library science" is intended to mean the <u>science</u> of contemporary librarianship. To use the term "science" implies the existence of a kind of theoretical and structured knowledge developed by the systematic use of scientific methodology. Clearly, no such substantial knowledge of information transfer and retrieval principles exists in this field. If, however, "library science" is intended to express the process by which we can create, test, and evaluate basic principles to guide and justify optimal procedures for information identification, transfer, and retrieval, then the use of that expression is not outrageous. This process is being followed by many workers in this field. The term "librarianship," on the other hand, is certainly not pretentious, but it does suggest an emphasis in this field on a practical knowledge of traditional library operations and the development of the "skills" of being a traditional librarian. Thus for some, "librarianship" connotes the skills required by one who deals with the traditional library related problems of how to organize books and bookshelves, reference and circulation desks, etc. On the other hand, "library science" connotes more basic generality. It suggests a basic concern with information, as opposed to books, and a concern with the problems and principles of <u>information</u> storage and retrieval: problems of how to organize, identify, store, search for, retrieve, and disseminate <u>information</u>. In what follows we hope to avoid any terminological dispute by using the expressions "library science" and "librarianship" interchangeably; both are intended to connote the theory and those procedures involved in activities of information storage and transfer.

We can now return to our first question and ask, "What is happening in contemporary librarianship?" One of our purposes is to indicate where we stand today in this field, and why the early high hopes and expectations for fully automated information retrieval systems are still "in the future."

Also, we are concerned with finding what the prospects are for having general principles to guide in the development of such fully automated systems. All of this, of course, is related to the question of what would constitute a relevant education in this field.

For librarianship, those years just after 1945 can be seen as a point of transition, certainly in terms of hopes and expectations. At the close of World War II, a number of developments emerged which were thought to have a fundamental impact on science, technology, and society. These developments were of special interest in the library field because they concerned theories and devices for dealing with information. In 1948 Norbert Wiener published his book Cybernetics. The thesis was that a new science called "cybernetics" was emerging. Cybernetics, said Wiener, was to denote the science of information processing and control for both biological systems and machines. The basic concept in the new science of cybernetics was information, and central to Wiener's thesis was the notion that various properties of the commodity of information could actually be measured in precise ways. Wiener went on to argue that the brain, which, in biological systems, is the central organ for information storage, processing, and control, could be analyzed in a completely mechanistic way. He further argued that the concepts of cybernetics could assist in understanding how intelligent, problem-solving behavior is related to basic mechanical information processing in the brain. All of these notions were of special relevance to library scientists because central to brain organization and operation is its library function; i.e., the brain and nervous system together embody a powerful system for information storage, search, and retrieval. Thus Cybernetics led some people to hope and expect that fresh new insights into biological mechanisms for information storage and retrieval would lead to new techniques for using computers to identify, store, and retrieve information in a mechanized library system.

Within a year of the publication of Wiener's book, another highly technical book which dealt with the concept of information exploded into print, creating still another wave of intellectual excitement. This was Claude Shannon's Mathematical Theory of Communication. If information was, in fact, the key concept of a new science of cybernetics, then Shannon's book was very important because it offered a precise theory to explain the meaning of, and how to measure, the amount of information conveyed by a message. His book led some readers to hope and expect that its ideas might be developed further so that we might have a complete theory of language and of meaning. With these concepts then, one could go on to develop a full theory of information search and retrieval.

Another development causing great excitement during those years was the electronic digital computer, a machine for storing and processing information, a general purpose information machine that could do any task that its programmers could precisely describe. Of course, by our present standards, those first computers were outrageously slow, costly, unreliable, and unwieldy to program and operate, but some people nevertheless were able to see the great potential of these machines.

Thus, there were early predictions about using a digital computer as an automatic language translating device. And, more relevant to our present discussion, there were early hopes about using the computer as part of an automatic library system - a system for searching and retrieving information at electronic speeds. There were predictions for getting machines to handle language in comprehending-like ways, and for designing machines that could store and analyze tens of millions of factual sentences and then answer questions based on that stored information. In fact, some predicted that future machines, properly programmed and given sufficient storage capacity, would be able to deal with ordinary language in such intelligent ways that a person who could only observe their output would think that he was observing the linguistic behavior of an intelligent human being.

Needless to say, many of these predictions now seem wildly extravagant, even irresponsible. Today, we are far from having such information systems, and one of the questions we want to consider in the pages that follow is why there is such a gap between the early expectations and the current accomplishments in this field. Why has progress been so slow, and what is it now reasonable to expect at the end of this decade? It will turn out that there has been a monumental misunderstanding, on the part of some, of what is involved in getting a mechanical system to deal with language in understanding-like ways. There are deep conceptual problems that were glossed over, and the true complexity of the problems that need to be solved require basic work at the very foundations of this field, rather than a mere continuation of experimental testing of new "tricks" for getting a computer to help with the problems of information analysis and retrieval. We believe that work at the foundation of library science must be coextensive with work at the foundations of information science. We will say more about this overlap subsequently, but first what has been happening in the field of library science? What kinds of information processing techniques have been developed?

2.2 Information Processing and Library Science: Some Distinctions

2.2.1 Control vs. Access

At the outset, we want to make some distinctions between different kinds of library information processes. The first is the most basic: it is the distinction between information processing for the purpose of access, and information processing for the purpose of control. What exactly is the distinction between access and control? The major purpose of a library (or information center, data bank, etc.) is to acquire, identify, organize, and store information so that the information will be accessible on demand, for use by its patrons. Thus, the major purpose of a library is to provide effective access to information, regardless of the exact type of information, type of patron, or type of query. And, needless to say, providing effective access to information, there are many problems of indexing, searching, relating, etc., to confront. In order to organize a library for the purpose of providing access, it is necessary to monitor and keep track of how the system is functioning, and this is what we refer to as the problem of control. To acquire, identify, store, retrieve, circulate, record, and, in general, run a library, somehow there must be information processing to control what

is going on. In most large conventional libraries, much time, energy, and information processing capacity must be devoted to the activities of keeping track of the books ordered, payments made, items received and those not yet received, books circulated and those overdue, the serials received and those in need of claiming, the books lost, and those at the bindery, etc. All of these activities having to do with keeping track and of monitoring the functioning of the system are what we call control functions. By and large, the two activities of information processing in libraries for purposes of control, and information processing for purposes of access, are logically separate and distinct.

The category of information processing for the purpose of control might be called "library systems analysis and mechanization of clerical functions." However it is called, we can further distinguish two aspects of the work; viz., practical or applied systems analysis on the one hand, and theoretical aspects of library systems design on the other. What do we mean by this split (which, incidentally, is not always sharp and exclusive)? For education in librarianship, the practical side of systems analysis and library mechanization is concerned with teaching the "how to do" those types of systems analysis, design, test, and evaluation tasks that seem to be necessary steps toward the mechanization of clerical (control) operations in libraries.

The teaching of "how to do" these tasks might be done via real case studies, or by simple artificial examples. It is not at all clear what aspects of the practical side of library systems analysis (and clerical mechanization) should be taught in any library school. And if it were to be taught, what is the proper kind of prior course work to serve as prerequisites. For example, how much systems analysis and computer programming should be taught in library school; or should all of this material be taught outside of the school (i.e., in the departments of Industrial Engineering, Computer Science, etc.). If a student does get some brief exposure to these subjects in library school, would he (or she) know enough upon graduation to go into a library and make a useful contribution toward mechanization of some of the control functions in that library? There seems to be a real need for skilled, experienced systems engineers who can do practical work in this area, but it is not at all clear how to train and prepare such people. Perhaps all of this kind of "practical" education belongs "on the job" and not in the classroom of a library school.

The theoretical aspects of this category, which covers the mechanization of control functions, is concerned with a miscellaneous host of questions such as, "What are optimal ways to encode and store bibliographic data, so as to minimize error rate?" The theoretical problems, in general, concern theories and models for how to perform some part of the control process in an optimal way, relative to certain constraints.

Who should be developing these theoretical techniques? Again, it is not at all clear that this kind of theoretical work falls into library science proper. Perhaps, as techniques, the work would more properly fall into the disciplines of Operations Research, Industrial Engineering, or Computer Science. The librarian, as such, is concerned with the use of such

techniques, but again the development and refinement of the techniques as such would seem to belong to a separate discipline. So much for questions of control. The major practical and theoretical aspect of librarianship, qua librarianship, concerns the problems and processes of access, not control. Let us now turn to this central problem.

2.2.2 Question-Answering vs. Literature Searching Systems

We made the distinction between information processing in libraries for purposes of control and information processing for access. We now turn to the problems of access and ask how the computer might be used to assist in automating aspects of the key processes of interrogation, search, and retrieval. First of all, exactly what is meant by the term "access"? Are there different kinds of information access? And is the process of obtaining access composed of sub-processes that can be described precisely? The general problem of access can be described as follows: a person wants information of some type or variety, for some purpose. Thus, the problem begins with a person who has an information need - which is a psychological entity not directly accessible to the library system. The library system has stored a wide variety of information "packages." The problem is to decide which items of information, if given to the requesting patron, would best satisfy his need for information. How might a computer be used to mechanize the search for so-called relevant information? This is the access problem.

It is now standard to distinguish between two rather different types of information needs which in turn are reflected in two rather different kinds of requests for information. On the one hand, a library patron might be seeking the answer to a rather specific kind of a question, such as "When was Isaac Newton born?" The desired answer is simply the birth date of Newton. This class of information access, where a person is seeking to obtain a specific item of data, is called data retrieval. Mechanized systems for providing access to this type of specific reference question are often called "question-answering systems." The distinction we want to make is the now standard distinction between two types of information access and retrieval systems: question-answering systems (aimed at providing specific answers to specific questions), and literature searching systems (aimed at providing relevant-useful literature in response to a request for information on some given subject or topic). Incidentally, this distinction between question-answering systems and literature searching systems is reflected in traditional librarianship, where problems and techniques relating to the former are called "Reference Studies" and where problems of the latter go under the heading "Cataloging and Bibliographic Organization."

2.2.3 Problems of Question-Answering Systems

What is involved in the design of a question-answering system, and in what ways might a computer be used in such systems? Before saying where the problems lie, we must indicate that there can be a rather wide spectrum of question-answering systems, ranging from rather simple, so-called "look-up systems," to very complex systems that deal with ordinary language in

comprehending-like ways. The simple systems already exist and are finding application in growing numbers. The more complex question-answering systems are still the subject of study and investigation, and where they are now operational, it is primarily for the purposes of study and research.

As an example of a logically simple kind of question-answering, consider systems that are used by most airline companies to keep track of seating on flights. The kind of information that is stored is both very limited and very highly structured. The type of inquiry that may be made (e.g., whether or not seating is available on a given flight on a given day) is very limited. From a logical point of view, this type of information retrieval system is simple, but very useful in those situations where the data in question is changing so rapidly that it cannot be put into book form because by the time it were printed, it would be out of date.

In a more complex type of question-answering system, the process of responding to a query would involve more than mere look-up of well structured data in some file. In the more complex case, the system may have to analyze linguistically a large amount of its stored textual data in order to logically derive the desired answer. That is to say, in those cases where the desired answer to a given query is not stored explicitly, the system must be designed so that it can deduce the answer (according to principles of logical deduction) if it is a logical consequence of some of the explicitly stored data. The designer of such systems is faced with the extraordinarily difficult problem of providing suitable rules of logic and deduction for the machine, so that it can deduce and thus make explicit the data and consequences that are only implicit in the stored data.

A related problem in the design of question-answering systems concerns the role of ordinary (natural) language in such systems. For example, if rigid rules of logic and deduction are needed as described above, the data must be represented in the machine in terms of some rigorous logical language (because it is only for such precise languages that rules of deduction now exist). Further, if the machine is processing information that is presented in terms of a precise logical language, there must be prior provisions for mapping into and out of this language and into and out of natural language. We can see that the whole problem of how to analyze ordinary language so that this mapping can be effected is, therefore, part of the larger problem of how to design really effective question-answering systems. Again, complex question-answering systems are still a subject of serious study. Simple systems already exist and are growing in members in all facets of our society. What about literature searching systems?

2.3 Literature Searching Systems

2.3.1 The Problem of Literature Searching

The problem of literature searching starts with a person (e.g., the library patron) who wants to locate literature (information) on some given topic or subject. Unlike the patron who might approach a question-answering system, he is not looking for the specific answer to a specific question,

but rather information about some subject. The problem of the literature searching system is to acquire, identify, and store incoming documents and to analyze a topic request in order to predict and then retrieve all and only those items of its stored documents that would most probably satisfy the information need of the inquiring patron. We might characterize the problem of literature searching most generally as a problem of inference and prediction in the following sense: the problem of the system is to be able to predict correctly (and then retrieve) all and only those of its stored documents that, when subsequently read by the patron in question, will satisfy his initial need for topic information. In order to be able to make such a prediction with any degree of correctness, the system must have at its disposal sufficient information, not only about the contents of its stored documents, but also about the patron whose information need motivates the entire search procedure. Thus the system must have clues by which it can identify documents and identify information needs, and it must have inference-making rules so that given this data it can predict about which documents would most probably satisfy a patron.

If this field were further developed in the direction of having some kind of an underlying theory of literature searching, we would know what kinds of clues and data a system would need in order to do the kind of prediction described above. This would be rather complex because ultimately, such a theory would have to deal with such concepts as information need, states of knowledge, content of a document, and finally, what it would mean to gain knowledge from a document and thus remove an information need. We are far from having such a theory, and yet, even without a theoretical guide for the system designer, there are a number of steps toward the design of mechanized literature search systems that can (and have) been taken. We have progressed since 1946 (when the first electronic literature searching system was constructed) by taking a series of small steps, one at a time, in the attempt to build better mechanized literature search systems. We take a few steps, test to see whether these have resulted in an improvement, and if so, search for the next steps to take. These so-called "steps" are techniques for obtaining access to stored literature, and they are also formal search rules for use by a computer as part of automating information retrieval.

2.3.2 Models of Literature Searching Systems

A model of a literature searching system is a precise description of the procedure for requesting, searching, and retrieving stored documents. In order to mechanize the procedure and thus step toward automated literature searching systems, the description (model) must be clear, complete, and precise enough so that at least the search aspects can be implemented by a computer. In the process of formulating a precise description of the search procedure, we must make a number of important simplifying assumptions about the problem of literature searching. As we move toward more complex and realistic models of the problem of literature searching, some of the simplifying assumptions are modified and, hopefully, made more realistic.

The sequence of key steps involved in a system for literature searching follow: (1) incoming documents are analyzed and identified for the purpose

of subsequent retrieval; (2) the identification of content is represented by assigning so-called index terms to every document; (3) the index information (plus associated bibliographical information about each document) is coded and stored for later search.

The problem of a literature search starts with a patron who has some need for information, and this need gets expressed initially as an informal request for information. In order to interact with a mechanical system, the informal request must get "translated" or "mapped" with a formal query. The formal query is a formulation of the request in the language of the retrieval system; viz., in the language of index terms. The index terms, in a sense, represent the common language that is used to bridge the gap between the documents and the request. Given that the documents are identified by means of index terms, and given that the patron's request for information is represented (as a formal query) by means of index terms, the problem of search is now reduced to the problem of how to operate on these two entities. We now describe several classes of retrieval models which represent ways of using a computer, given a query, to predict which documents will most probably satisfy the inquiring patron.

2.3.2.1 The First Class of Models

The first and simplest model in this first category consists of the following elements: documents are identified by assigning to each, one or several index terms, and every query consists of a single index term. The search procedure consists of selecting and retrieving only those documents which have the query term among its set of index terms.

The second model in this category is called the overlap model. In this case the documents are identified by assigning one or a set of index terms to each. The query consists of one or a set of index terms. The search procedure consists of selecting and retrieving all and only those documents whose index terms overlap those in the query set at a specified threshold. For example, if the query consisted of say four terms, the search rule might specify an overlap of three or greater, thus retrieving all documents which had at least three of the query terms assigned to it.

The third in this sequence of first models is the Boolean model. The indexing of documents remains the same as the cases described above, but here a query consists of a set of index terms connected by any combination of truth functional connectives to form a Boolean string of index terms. The search procedure consists of selecting all and only those documents whose set of index terms are included logically in the set described by the Boolean query (i.e., those documents whose index sets imply those of the query).

In all three of the above models, it is assumed both that every document is either relevant or not relevant to a user's need, and that the problem of the literature searching system is to predict which documents are relevant and to retrieve them. Thus the system makes a binary (two-valued) decision for each document relative to each query: it either retrieves it or not, depending on the retrieval rule. The output is the set of retrieved docu-

ments. There is no attempt to rank the retrieved documents.

2.3.2.2 The Second Class of Models

Now consider those cases not where the retrieval rule divides the collection into two disjointed sets (relevant or not), but rather where there is a degree of match computed, and thus the output of a search is a list (or actual set) of documents (from the stored collection) ranked by degrees of computed relevance. In order to describe this class of models, we first must introduce the concept of closeness or similarity. Closeness, of course, is a key concept in these models of retrieval systems, because here we move from binary rules for deciding which documents to select for retrieval, to search rules based on degrees of closeness. We will consider closeness computed between the following entities: document indexes and queries; index terms and other index terms; and document representations (vectors) and other document vectors.

One can interpret a set of index terms (assigned to a document) as a vector in an n-dimensional space, where n is the total number of different index terms. This vector can be thought of as identifying and representing the document in question. The orientation of the vector identifies the location of the corresponding document in this n-dimensional space. If a query is represented similarly as the set of index terms expressing the user's request for information, then one can measure the angle between the query vector and document vectors. This is one of many different ways of measuring the closeness between document representations. The point here is merely to indicate that if documents and queries are represented as vectors, we can formulate a retrieval rule that measures the closeness between the query and all document vectors, and thus the system can rank (order) the collection by degrees of computed relevance. This type of retrieval model might be extended further by moving from binary indexing (where each index term is either assigned to a document or not) to a weighted indexing (where index terms are assigned to documents with a weight indicating the degree to which that index term applies*). Weights might also be assigned in query terms, thus permitting the retrieval system to compute a measure of closeness between pairs of weighted vectors.

Among the set of models in this second category, we include those that employ various forms of associative searching techniques. There are two forms of associative searching: associative searching in "index space," and associative searching in "document space." Associative searching in index space is a technique for taking a given request and expanding it by adding to it (disjunctively) those other index terms that are computed to be closest to the given ones. This means that given any terms, say I_j, that might appear in a query, the system can enlarge that query to include other terms close to I_j.

*A more precise formulation interprets the weight of an index tag I_j relative to a given document D_i as an estimate of the probability that if a user were to be interested in D_i, he would be searching for that kind of information under heading I_j.

Once one, or a set of documents is selected by a retrieval rule, that retrieved set can be enlarged to include other documents that are computed to be close to the initially retrieved documents. This technique, called associative searching in document space, uses various statistical measures of closeness to compute closeness in document space by selecting documents whose index set is "similar" to those in the initially retrieved set. Mathematical measures for computing degrees of closeness (or similarity) are an important tool in the field of information retrieval. Measures of association have been studied by a number of workers in the field of information retrieval, and at least 15 different measures have been proposed.* Different measures behave differently in selecting terms (or documents) close to a given one. There are both many different stages in a search (ranging from the early stages where the searcher knows very little about how well his request has been formulated or what kinds of documents are available, to late stages in a search when this kind of information has become available) and different kinds of information needs (ranging from emphasis on Precision at the expense of Recall to emphasis on the reverse). The study of associative searching is an important part of the larger problem of information searching, and it is a problem area that we chose to emphasize in some detail in the design of the Information Processing Laboratory. A detailed discussion of how the Laboratory is used to teach the techniques of associative searching is contained in LABSEARCH,** which is one of the volumes that make up this Final Report.

2.4 Intellectual Access: A Closer Look

2.4.1 Initial Remarks

What kind of a problem is the information retrieval problem? Roughly speaking, the problem is that of how to obtain access to all and only those items of information which, when read by the patron in question, best will satisfy his need for information. In order to have an optimal (or near optimal) solution to this problem, we need a theory of information search and retrieval. What would be involved in the construction of such a theory - what fundamental concepts would have to be explicated, and what types of relationships between fundamental concepts would have to be constructed? When might we reasonably expect to have a complete theory for the problem of intellectual access?

If it turns out that a theory of information transfer and retrieval is so subtle and complex that we cannot realistically hope to have it "all together" for, say, another decade, what then? If our primary concern is the design and development of really effective information retrieval systems, do we in fact need to wait for a complete theory of intellectual access to emerge? Perhaps it is possible to design really effective systems without having a theory of such systems. This seems to be an important point at which to probe. Thus we ask, as we did above, what kind of problems are logically similar to it? How might we learn the best ways to attack the problems of access with-

*See, for example, J.L. Kuhns, "The Continuum of Coefficients of Association," Statistical Association Methods for Mechanized Documentation, National Bureau of Standards, Miscellaneous Publication 269, Washington, D.C., 1964.

**Mignon, Edmond and Irene L. Travis, LABSEARCH: ILR Associative Search System Terminal Users' Manual, Berkeley: Institute of Library Research, University of California, September 1971.

out the benefit of having a theory? We will characterize literature searching as a special kind of problem solving activity. Furthermore, we will suggest how learning about and teaching this type of problem solving can be approached via an on-line Information Processing Laboratory.

2.4.2 Another Look at the Problem of Access

What would it mean to have solved the literature searching problem? At one (deep) level, it would mean having a complete theory of how to search and obtain optimal access. What would be involved as component parts of such a theory? If such a theory were to be cast in operational terms, it would have to provide rules describing: how to process a request for information, and how to analyze a set of documents relative to a given request in order to select (and then rank) those documents which, when read by the requesting patron, would satisfy his need for information. This would imply, among other things, that there be rules for predicting how a document will be comprehended, and how comprehending that document will alter the state of knowledge (or belief) of the reader, and thus how that document will tend to satisfy the original, so-called, information need of the patron in search of information. If, as we are here suggesting, an operational theory would have to include rules for predicting the impact of reading a document on the mind of its reader, we should see immediately that any complete theory of information search and retrieval would have to be extraordinarily complex since, in some sense, it would have to include a theory of comprehension. By a theory of comprehension, we mean a theory of how text, when read, affects the information (and belief) states of its reader. This in turn would have to presuppose that we have some well formulated mechanical theory, or model, of mind, and how information is accepted by and subsequently modifies a mind. Furthermore, in order to talk about how information changes what an intelligent receiver knows (or believes), we need to explicate further the meaning of knowing (and believing).

Well, as you can see from even these few remarks, any complete theory of information transfer and retrieval presupposes a prior theory of information formulated as part of a larger theory of knowledge, comprehension, intelligence, and behavior. Surely such "prior" theories would be much more complex than any physical theories that we have today in any field of contemporary science. We are suggesting that any fairly complete theory of intellectual access presupposes a theory of information and knowing and this means that library science, at its very foundations, merges in some aspects with a new science of information and knowing. Such a theory of information and knowing would be so complex that it would be most unreasonable to expect to have it at hand within a decade (or even a century). If this is the case, what are we (i.e., the information system designers) to do? Does it mean that we must sit and wait for a complete theory before we can proceed to develop improved retrieval systems? Or is it possible to construct better retrieval systems without the benefit of a full theory? And, if we accept the latter (as perhaps we must), then how do we proceed? It turns out, of course, that in other areas we don't need a complete theory in order to develop useful, effective systems. And we believe that the same holds both for this field and for the development of improved literature searching systems.

2.4.3 Literature Searching as Complex Problem Solving

What constitutes a problem, what represents the solution to a problem, what are ways of finding solutions to a problem, and, of special importance to educators, what are ways of teaching and learning how to find solutions to a problem? The problem that we shall be considering is the problem of literature searching, and our primary concern will be in inquiring how one can learn how to find good solutions; i.e., how to retrieve effectively relative to a request for information.

In order to clarify literature searching as a kind of problem solving activity, we first should consider the nature of problem solving in general. However, we cannot because the subject is much too broad and complex. Instead consider certain rather formal types of problems as represented, for example, by board games. As a specific example, consider the game of chess. For chess, a problem would be how, say, White could go from a given board configuration to a subsequent board configuration in say four moves independent of any moves that Black might make. One might think of both this kind of problem, and the transition from problem to solution, in a geometrical way - as moving from one point to another in a complex space. The initial board configuration is represented by a point A in a maze; the desired board configuration is represented by a different point B; and the problem, of course, is how to make the proper moves in order to go from A to B in a fixed number of steps. In the case of chess, each player is faced with a set of possible (legal) moves at each step (turn), and he must select the best in order to go from A to B.

If we were to consider, say, theorem proving in logic instead of the game of chess, a problem would be to find a proof for a given theorem. The initial set of axioms (of logic) would correspond to the initial state, the theorem to be proven would correspond to the desired state, and the solution would consist in finding a sequence of transformation rules that when applied at each step, would allow the theorem in question to be derived from the initial axioms. Here again we can think of problem solving as moving (by selecting one out of a set of possible moves) from initial point A to the desired state B in a maze. The notion here again is that, geometrically speaking, a problem can be thought of as a gap between two points A (the given state) and B (the desired state). The solution consists in finding how to move from A to B; i.e., how to find a sequence of moves which, when connected, form a chain from A to B. The links in such a chain are selections made from the set of possible (legal) moves of the game - whether it be chess, theorem proving, or, as we shall see, literature searching. If we consider literature searching as a type of problem solving, what corresponds to the initial state (that we have called A), what corresponds to the desired state (called B), and, most importantly, what corresponds to the set of allowable (legal) moves from which the chain is constructed connecting A and B? We can consider a problem originating when a patron's mind is in the state of having an information need. We can consider the solution consisting of the patron's subsequent state of mind after having digested the desired relevant documents. However, for our purposes we can deal not with states of mind, but rather the formal request (as representing the initial information need) and the desired (relevant) documents (as representing the change to be made, when read, in the patron's state of mind). Thus the formal request corresponds to our point A, and the set of "relevant" documents corresponds to the point B. Now, what corresponds to the set of legal moves in this game of literature

searching? We suggest that the class of formal techniques for intellectual access (described in Section 2.3) corresponds to the set of allowable moves. Thus the problem of intellectual access to stored literature, as we are here portraying it, is the problem of how to go from an initial request (for information) to a set of so-called relevant documents, by constructing a chain of formal techniques of access which, when applied, lead from the request to the desired output documents. Before we elaborate on this notion, consider first what it means to speak as we did earlier of a complete operational theory for some problem situations.

To say that we have a complete operational theory of chess would mean that given any board configuration A, and any desired configuration B, the theory could specify a set of moves which, when applied, would lead from A to B. In certain games such as NIM, there exists a complete theory, and thus with it one can always guarantee not to lose at the game of NIM. In logic what we are calling a general theory would correspond to a decision procedure, i.e., an algorithm that guarantees an answer to the question "Is T a theorem?" If we have a general theory, then there is no problem. We simply apply it and find our solution. If there were a general theory to the problem of literature searching, it would prescribe exactly how to operate both on a request and on the set of stored documents in order to select all and only those that would satisfy the patron's information need. Of course, however, no such general theory exists.

Now whether in chess, logic, or literature searching, how does one solve problems without the benefit of a general theory? Simply stated, a problem is solved by making a move sequentially, looking at the consequences (perhaps in terms of how close it has moved us in the direction of the desired solution), and then selecting the next move, until (if possible) we have moved from A to B. That is, problems are solved by learning how to formulate, execute, and improve problem solving tactics and strategies. In the case of literature searching, what exactly does all of this mean?

2.4.4 Literature Search Tactics

In order to clarify the notion that the process of literature searching is a problem solving activity (similar to chess, although, of course, not played against a rational opponent), consider an on-line interrogation, search, and retrieval system that would function as follows: the documents of the collection consisting of professional journal articles are stored along with complete bibliographical records for each, including titles, authors, index set for each, abstracts, and list of papers that each cites. A patron approaching the system in search of articles that would satisfy his information need would first have to formulate a formal request. This means that he must select a small set of just those index terms that he thinks would best capture the desired documents, and, as part of the request formulation, he would have to connect the chosen index terms with the appropriate combination of logical connectives ("and," "or," "not"). We might think of this as his opening move. There are very many possibilities, and the patron must select that one (initially) that he thinks will be a good one. Given his initial request,

the system might first respond by telling the patron how many documents there are that satisfy the given request. This response by the system confronts the patron with his next "move." Here again, there are a large number of possible (legal) moves, and it is up to him to select one that he thinks will be good. If the system has told him that his initial request will cause a large number of items to be retrieved, he may modify it in a number of ways in order to narrow it. He might, for example, select some different terms, assign differing weights to his request terms, or modify the logical structure (as opposed to the content) of the request. Or he may decide to do nothing and thereby retrieve all the items specified by the original request. If he does not wish to modify his initial request immediately, he might request that the system display the titles and indexes of those items specified by the request. When the system responds, the patron has some concrete feedback indicating, in a sense, where in document space his original request has "landed" him. From the bibliographical records he can tell to some extent whether or not his request in fact is leading to documents that will be "relevant." He might now narrow his original request to eliminate those items that appear less useful. The system, in turn, responds again by indicating the number of documents specified by the modified request. This number may be too small for the patron. He wants to expand the search in a slightly different direction. He decides to use associative searching in "index space" (see Section 2.3). Now he must select one of a large number of measures of statistical closeness between index terms. This is his next move. He might have decided to expand his search in document space instead. The result of these moves is a list of items ranked by some computed measure of relevance. The patron now either must trim the list, or expand it using a different search technique, etc. This process of selecting a move (from a large number of possible techniques for access) continues until, hopefully, there is convergence; i.e., until a set of documents which satisfies the patron in question is finally retrieved.

We have sketched a process of intellectual access that we feel corresponds to the process of problem solving in other areas. Now we want to raise some questions about how this type of problem solving activity can be learned, and how it can be taught. Again, searching is at the very heart of the so-called library problem. It is a complex type of process. How can we learn to perform it in effective ways? And how can we teach students good tactics and strategies for this complex "game" of literature searching?

2.4.5 Learning about Literature Searching

We described the process of information searching as a game involving moves, and it is clear that sequences of moves with certain purposes correspond to search tactics. Some people are very effective in their information searching activities because they possess intuitively good search tactics and strategies. However, if they are using conventional library facilities, the range of their "moves" is extremely limited; e.g., there is no possibility of using weights in a request, etc. How can we learn and teach how to conduct effective searching, not only to teach the meaning of the variety of formal techniques (the moves), but also to learn about and teach tactics and search strategies?

Begin by returning to the basic elements of a search strategy, i.e., the individual moves, or individual techniques of access. What exactly are these techniques, and how can they be learned? We argue that these techniques for access are, in fact, types of logical tasks; i.e., each technique is a different type of tool for performing logical work on a body of stored information. And how does one learn about a tool - about how it works and how it can be misapplied? We learn about tools by using them! We cannot best learn about tools (whether physical or logical) by being told or by having them described, any more than we can learn about how to ride a bike or to swim by being told. We learn about tools both by using them in a variety of circumstances, and by seeing how they work and fail to work when applied in different situations. And in the case of information retrieval techniques, we can learn best about their effects by testing them on different types of data bases: by using them to search and by looking at the retrieval consequences of that use.

2.4.6 The Information Processing Laboratory

By our interpretation, the Information Processing Laboratory, from its inception, has been one where students could sit at a remote terminal on an individual basis and interrogate, search, and analyze bibliographical materials stored in a central digital computer. The idea was that a variety of corpora of bibliographical data would be stored, each perhaps indexed in a different way. There would be a wide selection of formal techniques for access, each callable from the terminals. Thus a student could select an access technique to be studied, and he could "call" it and "exercise" it on one or more of the stored data bases. Having both stored data files and search and display commands that can be activated from the terminals allows students to test and examine quickly the retrieval consequences of using these techniques singularly and in combinations. Thus, the central purpose of an Information Processing Laboratory is to provide for a level of depth and understanding of a very complex set of search procedures. This kind of understanding can come only with the use of a computer which can derive and display the consequences of using complex rules. This type of understanding of logical techniques at the very core of librarianship cannot come from lectures alone, but rather must come via a first person interaction. We have suggested that the problem of obtaining deep intellectual access to stored literature in a library system is a special form of problem solving, and, furthermore, that there is a similarity in the problem solving activities of literature searching and theorem proving in logic. A person cannot learn how to prove theorems merely by being told. He has to immerse himself in this type of problem solving activity and begin to experience the different kinds of available clues and how the use of these clues can lead him even closer to the desired solution. Theorem proving, like chess, has certain rules which describe those "moves" that are legal. But these (transformation) rules indicate merely what is allowable, not which ones are most suitable to a particular stage of an attempted proof.

In the case of literature searching, there is a wide variety of allowable (legal) search moves that can be made from the time that a search is initiated with the selection of index terms, to the time the search is completed with

the selection of those documents that appear to satisfy the initial request. The tactics involving the selection and use of these allowable search moves determine the direction and outcome of the search. Thus, in order to understand when and how to expand a query, when and how to narrow the search formulation, and when and in what direction to go deeper, the user needs prior experience in actually using these techniques.

2.4.7 Summary Remarks

Education is preparation for the future. We cannot see the future. In a sense it is not "out there" to be seen: it is in the process of being created by what we do today. Nevertheless, we do conjecture that the digital computer will play an increasing role in mechanizing various aspects of information processing in libraries of the future. Thus, in thinking about and planning the Laboratory, we decided to emphasize teaching the use of the computer to assist with the problems of access, specifically the problem of accessing documents as opposed to accessing data. We have been designing a new kind of facility where advanced library students can learn about both the logic of literature searching and how to solve literature search problems. We have argued that literature searching is a complex type of problem solving activity. With the digital computer it is possible to devise and use a large class of different kinds of search techniques (e.g., different measures of closeness based upon measures of statistical association between index terms) used during a search. However, in order to learn about tactics and strategy of deep searching, one must use these techniques under a wide variety of conditions. Thus the primary purpose of the Information Processing Laboratory is to provide an educational and research environment to develop insights and skills needed to interrogate and search effectively. And also, such a Laboratory can function to stimulate, motivate, and prepare students for a future both where computer techniques for information search will be used more widely in many operating libraries, and, hopefully, where the computer will be more widely used in schools of library science as a new vehicle for learning about and teaching key aspects of the problems of access.

3. LABORATORY ORGANIZATION AND FACILITIES

3.1 Introduction and Summary

The basic organization of this report is designed to answer three questions: why is it important to have an Information Processing Laboratory; what equipment and facilities does our Laboratory have; and finally, how does the Laboratory function in an educational context. Chapter two has given our rationale for why the Laboratory should exist and continue, and chapter four will discuss how the Laboratory is used as an advanced educational resource in the School of Librarianship at U.C. Berkeley. In this current chapter we will attempt to simplify the transition between why and how by describing what the Information Processing Laboratory is in terms of hardware, program systems, and staff.

The computer equipment in the Information Processing Laboratory is an on-line video terminal system; the primary use of the Laboratory is for library education and research. This combination of on-line video system and a commitment to educational use defines in large part the special status of the Laboratory, and we will try to describe fully our on-line equipment and software system. The discussion of an on-line laboratory as an educational resource for teachers and students will be taken up as the main theme of chapter four and will not be covered here.

In this chapter we will also describe the staff organization of the Laboratory and the structure of student work sessions. The basis for our discussion of student usage will be the 1969/70 and 1970/71 academic years when the Laboratory was open and available for student use, and was a regularly scheduled component of several courses offered in the School of Librarianship at U.C. Berkeley. Though the basis of material presented is real and historical, we would still emphasize the provisional and prototypical nature of the data. A great deal of our work reported here represents first-cut approximations of what a laboratory facility could or indeed should be. We hope that this report can be used as a tool in the planning and design of other laboratories and similar research and educational facilities.

3.2 On-Line Terminal System

3.2.1 History of Laboratory On-Line System

From its inception, the Information Processing Laboratory has been conceived as an on-line facility for education and research in librarianship. On-line is a term used to mean that a set of keyboard and display terminals are in direct and continuous contact with a central computer system. The Laboratory's on-line system is also a time-sharing system, in which the central computer performs two or more tasks during the same time interval by interspersing processes, allocating small divisions of total time to each task in turn. The Laboratory system also involves teleprocessing which requires establishing a remote communication link

via telephone lines, between the central computer and the Laboratory's keyboard/display terminals.

From the user's point of view, the value of on-line systems lies in their accessibility and in their rapid response time. The keyboard/display terminal is conveniently accessible, and what is more important, because of time-sharing and teleprocessing, the response of the computer comes back in a matter of three to four seconds. For a facility which is committed to education in librarianship and information science, an on-line facility is the only way to provide both immediacy and direct experience with abstract material. This is especially true where the basic data files are textual rather than numeric.

Thus the development of a viable on-line environment is not merely a technological fad for the Information Processing Laboratory. Remote accessibility, time-shared independent terminal operations, and immediate response time are all important foundations of the Laboratory's educational philosophy. Initially, the first attempt at such an environment was built around mechanical terminals such as the Teletype Model 35 or the IBM 2740 remote terminal typewriter. During 1967 and early 1968 the first Laboratory programs were developed and run on these two mechanical terminals.

However, the display of text material on a typewriter device is slow and noisy and inhibits rapid scanning of data. For this use a CRT (Cathode Ray Tube) or video display terminal is a superior device. In the fall of 1968, the Institute of Library Research and the School of Librarianship requested funds from the University of California to purchase a system of cathode ray tube display/keyboard terminals to be used for educational innovation in librarianship. The argument in favor of CRT terminals was expressed as follows: "Because of the large amounts of data that are required for presentation to a user at a remote terminal, we now feel that a visual mode of output presentation on a cathode ray tube would be much more desirable than a typewriter. The use of a character-by-character printer is slow and costly, and it deprives the user of the ability to grasp a large amount of data in a single glance and select some small sub-portion of detailed use."

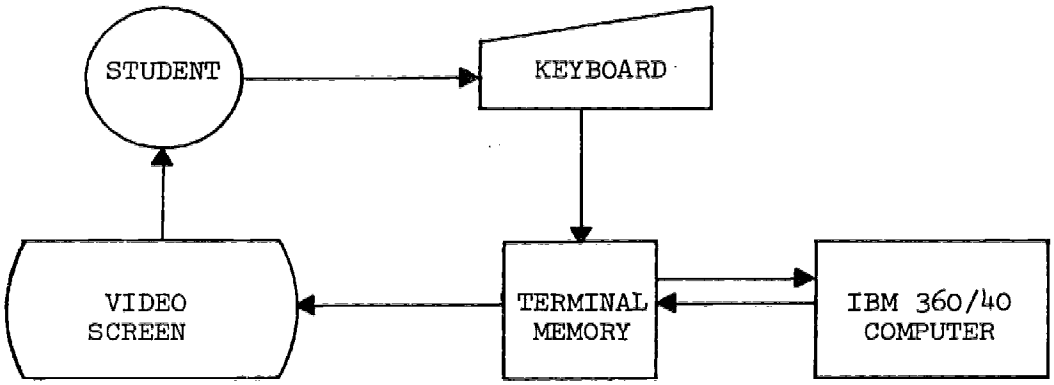
In the spring of 1968, the University generously approved the request and allocated money from a fund to support special innovative projects in instruction. The allocation was large enough to purchase three video display and keyboard input/output terminals, two memory control units, and two modulator/demodulator devices, to operate with two leased telephone lines between the School of Librarianship and the Campus Computing Center. The equipment which was purchased was a Sanders, Inc. Model 720 Communication System.

3.2.2 Terminal System Components

The purpose of the entire terminal system is to put the student-user into immediate contact with the Information Processing Laboratory data bases and processing programs which are stored in the IBM 360/40 computer system. The student formulates various program parameters and enters these data into the computer by using the terminal keyboard.

The program's response is displayed as a text message on the terminal's video screen. These two devices (keyboard input and video output) are rapid enough to provide an environment for real-time communication between the student and various bibliographic and information retrieval systems.

FIG. 1: SCHEMATIC REPRESENTATION OF THE COMPONENTS OF THE EDUCATIONAL TERMINAL/COMPUTER SYSTEM



The student formulates an input (request, command, interrogation) for a Laboratory system program. The formulation is entered character-by-character from the keyboard into the individual terminal memory, and from that memory back to the terminal screen. The student can review the formulation, correct errors, etc. When the formulation is ready for transmission, a special "SEND" key is operated and the contents of the terminal memory are relayed to the IBM 360/40 computer. (The message also continues to be displayed on the screen.)

A hierarchy of programs receive and process the student's formulation, perform the requested actions, and transmit a response back across the telephone circuit to the terminal memory. This is in turn converted into display characters, and the screen is changed to represent the program's output response to the original student input request. And so this input/output cycle continues. The user's input may consist of a request to sign on, to load a program, to search a file, to display a file, or to display a retrieval result. The program's response may be to carry out the requested action (e.g., to load a program or search a file), or merely to note that an action has been carried out (e.g., JOHN LOGGED IN, PLEASE SPECIFY PROGRAM).

The basic requirements for effective operation of an on-line educational facility can be summarized as:

- multiple terminals with both input (keyboard) and output (display) capabilities
- location and scheduling convenient for students

- independent and asynchronous individual terminal operations
- real-time response cycle
- legible noise-free display of text material
- minimum usage of computer time and storage
- stability and reliability of programs

All of these requirements are necessary to provide the environment for effective student use of an on-line facility. In the Information Processing Laboratory, these requirements are satisfied by a combination of the hardware system, terminal monitor program system, and the configuration of the central computer, the IBM 360/40.

The hardware system consists of two subsystems, one in the Information Processing Laboratory, and one in the Campus Computing Center. A single full duplex telephone line acts as a communication link between the two subsystems. In the Laboratory, the configuration consists of:

- modulator/demodulator (modem)
- memory control unit
- three terminals each of which contains
  - 12-inch video (CRT) display screen
  - 61-key input keyboard
  - 1,024-character memory

The modem conditions signals for transmission or reception across the telephone lines, thus allowing the entire subsystem to be distant from the central computing system. The memory control unit handles scheduling, routing, and queuing problems arising from the three Laboratory terminals, and maintains the independent and asynchronous operation of each individual terminal. Input and output are handled by the individual terminal keyboard and screen. Each screen is controlled by a separate 1,024-character memory. All screen transmissions, either from the keyboard or from the computer, are routed through the memory control unit.

(A detailed technical description of this particular terminal hardware system may be found in 720 Display System Reference Manual, Sanders Assoc., 1970. A summary of some of the major features of this description has also been given in Section 3.6.)

The availability of hardware does not in itself create an operating system, and the equipment in the Information Processing Laboratory was no exception to this rule. The goal of the Laboratory facility is to allow three students to use each of the three Sanders terminals simultaneously, with the options of each terminal operating a different program or all three terminals operating the same program independently and asynchronously. In order to meet this goal fully, it was necessary first to augment the core memory configuration of the IBM 360/40, and

second to develop an executive software system, in order to support realistically the kind of facility and operational usage which appeared to be desirable.

In order to augment the core of the 360/40, the combined financial support of the Information Processing Laboratory and the File Organization Project* was used to add an extra 128,000 characters of memory to the 360/40. This addition permitted the 360/40 to operate with multiple users in fixed-task partitions. This in turn allowed the Laboratory System to be available during large blocks of time without incurring high Central Processing Unit (CPU) time charges and without displacing other users. In a fixed partition multi-task system, machine users are charged for CPU time actually used rather than for the clock time which has elapsed while the user's job is on the machine. This is a logical arrangement because the computer's processing and input/output resources are being shared among many different users. The impact of this for Laboratory operations is favorable because CPU charges are incurred only when there is actual activity on the terminals. Thus, the time spent reading displays, entering inputs, copying results, etc., does not count as CPU time and consequently does not incur any charges. This reduces the CPU usage time costs to about 30% of the normal CPU hourly charge.

3.2.3 Terminal Monitor System

The largest of the 360/40 user fixed-task partitions is 108,000 characters. This area was allocated for the development of a Laboratory Terminal Monitor System and the operation of the Laboratory's bibliographic programs. Initially we hoped to find an extant software terminal monitor program to support our operation. Such a program would resemble many current time-sharing or teleprocessing applications. However, within the constraints of the 360/40 resources available, no such program was found, and thus again with joint support from the Laboratory and from the File Organization Project a Laboratory Terminal Monitor System (TMS) was designed and implemented by staff programmers of the Institute of Library Research. This proved to be a major undertaking both in cost and calendar time, although the results have been satisfying in terms of system performance and capability.

The major goal of TMS is to permit simultaneous operation of all three Sanders 720 Terminals. Two separate options are available. All three terminal users may operate the same program independently and asynchronously, although there is only one copy of the program loaded into the Laboratory partition. (This is made possible by requiring that all Laboratory bibliographic programs be "re-entrant," i.e., not contain any self-modifying instruction sequences.) A second option is for all three terminal users to request and operate simultaneously three

*The File Organization Project (OEG-1-7-071083-5068) is designed to study the problems of organizing and using large bibliographic computer files. The common interests of this project and of the Information Processing Laboratory have provided strong mutual benefits to each, especially in reinforcing the connections between education and research.

different Laboratory programs. The only limitation to this second option is that there be sufficient space available to meet each program's requirements.

For the programmer who is developing bibliographic application programs for the Laboratory, TMS provides a library of program routines to handle timing, screen display, disc access, input/output transmission and decoding, and hard-copy output. During the operation of an application program, TMS attempts to trap program loops and interrupts to avoid system crashes. In addition, various other on-line debugging aids are available.

For the terminal user, TMS provides a simple command language with which the user can sign on or off, and request that a specific program be loaded. This version of a command language is local to TMS and need not be used by the application programs; each of these programs establishes its own conventions for interacting with the program user.

When TMS first begins operation, the following pair of messages is directed to each terminal:

TMS IN OPERATION

WAITING FOR LOG-IN

This indicates that TMS is waiting for the user to identify himself at the terminal by typing in an identification code of up to four characters. This code is used by the system to identify hard-copy outputs, to name student data files which may be generated by application programs, and to record other terminal action patterns such as searching strategies, retrieval results, exercise results, etc.

After receiving a LOG-IN message, TMS displays a request to SPECIFY PROGRAM. The terminal user responds with the name of the program that he wishes to have loaded and assigned to his terminal. TMS then scans its catalog of Laboratory programs in order to locate and load what has been requested. There may be difficulty in loading the requested program. The TMS message PROGRAM NOT FOUND indicates that the name supplied is not the name of a program currently in the TMS library. Another possible TMS error message which may appear is: NOT ENOUGH CORE TO LOAD PROGRAM which indicates that insufficient core storage remains in the computer's Laboratory partition to load the requested program and/or its data areas. Either of the above messages is immediately followed by the message: SPECIFY PROGRAM which invites the user to try again. In the event that the requested program is successfully loaded, control is transferred directly, and the succeeding events are determined by the application program itself.

Depending upon the individual program and assuming that no errors have occurred in its operation, the terminal user will at some point exit from the user program and control will return to TMS. When this is done, the message: NORMAL EXIT FROM USER PROGRAM will appear followed by SPECIFY PROGRAM which allows the user either to specify a new program

to be loaded or to sign off. The process of leaving the system is called "logging out." When the TMS message SPECIFY PROGRAM appears, the terminal user may respond with the name LOGOUT. The system will respond with the messages:

XXXX LOGGED OUT

WAITING FOR LOGIN

which indicate that a new user may now identify himself to the system. If there are no new users for the terminal, the command DISCONNECT is entered, and the specific terminal is then effectively cut off from further communication or operation.

3.2.4 Inventory of Operating Programs

In our discussion of TMS, we mentioned bibliographic or application programs. In this subsection we will give a brief listing of all the programs currently cataloged and available in the Information Processing Laboratory System. The off-line support (e.g. for file generation) programs used by some of the on-line programs will not be listed separately. For each program cited we will give the following items of information: Name, Purpose, Data Base, Documentation, Programmer. Most, but not all of the programs cited will be discussed in detail in chapter four. All the Users' Guides mentioned as documentation are published as volumes of this current Information Processing Laboratory final report.

A. Name: BROWSER

Purpose: to provide a capability for examining index files used by the CIMARON System.

Data Base: index files used by CIMARON (see below); currently consists of Santa Cruz Author and Subject index files and San Diego Medical Society Author, Title and Subject index files.

Documentation: Chapter four of The CIMARON System: Modular Programs for the Organization and Search of Large Files by Arjun Aiyer.

Programmer: William Harrelson

B. Name: CIMARON

Purpose: search and retrieval operations processed against any MARC II structure data base.

Data Base: 95,000 monograph catalog records drawn from U.C. Santa Cruz library system; 5,000 monograph catalog records drawn from San Diego County Medical Society.

Documentation: The CIMARON System: Modular Programs for the Organization and Search of Large Files by Arjun Aiyer.

Programmer: Arjun Aiyer

C. Name: DISCUS

Purpose: compiler/executor for Machine Tutorial Mode (Computer Assisted Instruction) programs written in DISCUS language.

Data Base: two courses currently exist: 201X - Subject Cataloging and 201XL - Subject Cataloging Laboratory.

Documentation: DISCUS Interactive System Users' Manual by Steven S. Silver and Joseph C. Meredith.

Programmer: Steven S. Silver (System), Joseph C. Meredith (Courses 201X and 201XL), and Rod Randall (Support).

D. Name: DOLBYC

Purpose: demonstration of algorithm for reducing personal names to a canonical or phonemic form.

Data Base: None. Data is entered from the terminal.

Documentation: The Dolby algorithm is described in Appendix I of A Study of the Organization and Search of Bibliographic Holdings Records by Jay Cunningham, et al., Berkeley, 1969.

Programmer: Allan Humphrey

E. Name: LABSEARCH

Purpose: to implement concepts of associative search and retrieval.

Data Base: 485 abstracts of papers in the field of Information Science.

Documentation: LABSEARCH: ILR Associative Search System Terminal Users' Manual by Edmond Mignon and Irene L. Travis.

Programmer: C. Ravi (Search Program) and Rod Randall (File Generation).

F. Name: MAID

Purpose: to test and monitor indexor behavior with the use of index term co-occurrence tests.

Data Base: index term file of Psychological Abstracts (1969).

Documentation: "MAID", ILR Tech Memo.

Programmer: Steve Jacobs

G. Name: REFSEARCH

Purpose: to implement system of analyzing reference questions in terms of channels, qualifiers, and services.

Data Base: 160 reference work titles of the U.C. School of Librarianship practice collection.

Documentation: Reference Search System (REFSEARCH) Users' Manual by Joseph C. Meredith.

Programmer: Allan Humphrey

H. Name: SPECULOR

Purpose: to assist in programmer debugging by providing displays of core memory and allowing on-line modification of core memory contents.

Data Base: None

Documentation: "Speculor", ILR Tech Memo.

Programmer: Rod Randall

I. Name: TMS

Purpose: to provide teleprocessing capabilities for three Sanders 720 CRT Terminals in the Information Processing Laboratory

Data Base: None

Documentation: TMS: A Terminal Monitor System for Information Processing by Stephen F. Smith and William Harrelson.

Programmer: William Harrelson and Stephen F. Smith

J. Name: TIME

Purpose: to read and display computer clock.

Data Base: None

Documentation: None

Programmer: William Harrelson

3.3 Laboratory Operations

3.3.1 Interface with Campus Computing Center

In this section we propose to describe some of the arrangements and structures which proved to be necessary for the Laboratory's functioning as an educational facility for librarianship. Under this heading we will include: extended working arrangements with Campus Computer Center; Laboratory operations and usage patterns; and Laboratory staff organization.

The Laboratory Terminal system we have described to this point consists of:

- three CRT input/output terminals
- communications link to Campus Computer Center
- IBM 360/40 computer
- Terminal Monitor System (TMS)

These are the nuclear components of on-line facility. However, in order to operate such a facility, there are serious scheduling and integration problems with other Campus Computer Center activities which must be resolved.

For example, if the Laboratory Terminal Monitor System were to occupy the entire IBM 360/40 computer exclusively, then other 360 users effectively would be locked out during two or three hour Laboratory run periods. This is not a feasible situation since there is a sizeable campus community of 360 users including other projects of the Library and the Institute of Library Research. Furthermore, if TMS were to occupy the entire 360/40, then the hourly cost of Laboratory operations would be $90.00 per hour (the current hourly 360 CPU rate) regardless of the amount of time each student might spend in non-terminal work (e.g., copying material, keying input, reading messages, etc.). These two drawbacks (cost and tying up the entire computer) required that a different hardware configuration be found in order both to reduce running costs and to maximize the efficiency of Campus Computer Center operations.

Out of the many possible solutions to this problem, the one which was selected was to lease an additional 128,000 byte module of core

memory for the IBM 360/40. With this extended core memory (now totaling 256,000 bytes) the 360/40 is capable of supporting a mode of operations called Multiple Fixed Task (MFT). In this arrangement, the computer's memory is subdivided into a number of fixed-boundary partitions, and the Central Processor (CPU) and Input/Output (I/O) resources of the computer are shared among all the partitions which can now run independent jobs.

The resulting core map is shown schematically below:

IBM Operating System (36K)				
Laboratory Partition (108K)	User A (80K)	User B (20K)	User C (12K)	360/40 256K byte core memory

Thus, even when the TMS is loaded and the Laboratory is in operation, three other users can be accommodated at the same time. In this manner Laboratory operations do not disrupt regular Computer Center users. When the Laboratory is not in operation, the Laboratory partition is used by other Library and Institute staff programmers on a priority basis to provide more rapid job turn-around.

Under this arrangement, the pricing algorithm adopted by the Computer Center is on a time-used rather than space-occupied basis. This means that while the Terminal Monitor System is loaded but not actively used (i.e., during non-terminal activities) there is no associated cost. This reduces the effective computer time cost to about $25.00 per clock hour for all three terminals.

However, to this time-used price should be added an overhead cost represented by the lease price of the extend core (plus one-half of a 2314 disc which is used for data file storage). This cost is estimated at $36.00 per hour, (based on $5410 combined monthly core/disc lease cost ÷ 150 hours). The aggregate operating cost is, therefore, approximately $60.00 per Laboratory clock hour, or approximately two-thirds of the previous $90.00 per hour estimate. (It should be noted that Laboratory operations do not entirely support the lease of the core and disc, and that other Library and Institute projects bear a substantial cost load for this hardware facility.) Thus, if the Laboratory is scheduled such that all three terminals are in operation, the cost per terminal hour is approximately $20.00.

3.3.2 Scheduling and Usage

Given the environment just described in the preceding paragraphs, scheduling Laboratory operations is a matter of providing enough convenient hours for students while also allowing other Library and Institute staff programmers an adequate chance to use the Laboratory partition for research and development projects. This constraint

occurs simply because many staff programmers depend upon the size and priority status of this Laboratory partition. The solution is simply to run the Laboratory for specified short blocks (not exceeding two hours) of time, and allow one or two hours between such runs to allow other users a chance at the Laboratory partition.

For example, during a period of medium usage, a typical weekly schedule might consist of running the Laboratory from 12 P.M. to 2 P.M. and 4 P.M. to 6 P.M. daily. The later hours would be shared by both students and staff. If usage increased, an early morning hour (9 A.M. to 10 A.M.) could be added. Weekend runs (unscheduled and unsupervised) were also encouraged and occasionally used.

The following are some of the major variables which greatly affect the usage pattern of the Laboratory:

a. System stability. The ratio of down-time to scheduled operations. Down-time may be caused by failures of individual programs, Terminal Monitor System, or the IBM 360/40 (hardware or Operating System). At this point, after a considerable effort toward this goal, Laboratory stability is at the 90% level and is still rising.

b. Academic Calendar. Berkeley is on a quarter system; hence during a simple October-May academic year there are two quarter breaks, three registration periods, and three final exam periods. During each of these eight time periods, usually one to two weeks duration, Laboratory usage is practically nil. Thus, out of an eight month period, the Laboratory is on a limited schedule during nearly one-third of this time. The result is uneven scheduling during the entire period. There appears to be little escape from this dilemma, unless the University returns to a semester system, or unless there are more two-quarter sequence courses to provide greater continuity across quarters.

c. Ratio of student hours to Laboratory hours. This ratio depends first upon how many students use a terminal at the same time. We experimented with a scheme of Lab partners, in which two students jointly shared the use of a single terminal during a Lab hour. More recently we have assigned only one student per terminal. Both situations are feasible, depending upon the academic context and scheduling problems which may govern any particular case. A second influence upon the ratio has to do with the length and complexity of Laboratory assignments. There is a wide variation possible, varying from brief familiarization or orientation exercises, to sophisticated, open-ended research problems. Simple familiarization may take as little as 1/2 terminal hour per student. Simple exercises can usually be performed in a single terminal hour, though any extension such as an invitation to the student to explore a program or a problem on his own usually pushes the session to an additional

hour. More sophisticated programs, such as the Associative Search system, may require two terminal hours even for short controlled exercises.

There is also a large variation in the level of Laboratory usage depending upon the type of course which provides the context for Laboratory assignments. Survey or introductory courses may include Laboratory work as one component of many course assignments. In such courses, Laboratory assignments might constitute twenty per cent of the total academic work assignment. It should be noted that such courses also have heavy enrollments. Research seminars are the alternate context for student usage of Laboratory facilities. This population tends to be small, although the number of hours/student is usually high. Ideally, there should be a balance in Laboratory usage between introductory courses and research seminars. In actual fact this mixture is determined by individual faculty interests, viewpoint, and conviction that the Laboratory provides useful material for students. Here, as in other areas, we are still evolving toward a context which can be determined equally by curriculum considerations and faculty commitments.

3.3.3 Staff Requirements

In order to meet the demands imposed by its academic situation, the Laboratory staff has always consisted of people with strong teaching and/or research interests. The staff has, up to this time, included three distinct levels of responsibility:

a. project management

b. academic integration

c. laboratory assistance

Each of these levels forms a distinct and important part of the functioning of the Laboratory as an operating facility, both in its initial and continuing phases. In the following paragraphs we will attempt to describe what role has been played by each of these levels.

3.3.3.1 Project Management

Project management is especially crucial during the development phase of the project. The development of the software and hardware resources of this Laboratory has depended very heavily upon the coordinated efforts of several analysts and programmers, many of whom were working on different projects or even on different campuses. For example, TMS was jointly supported and developed by the Laboratory and File Organization projects. Goals, funds, staff, and hardware were all fully shared; similarly for the development of DISCUS, the Laboratory Computer Assisted Instruction (CAI) system.

The role of a project manager in this area was to coordinate diverse efforts, to monitor allocation of financial and human resources, and most difficult of all, to formulate and represent the Laboratory's

long-range goals to other technical staff within the development effort. In this role the project manager also provided feedback to the technical staff concerning program operation, the need for improvements, the success or failure of program modifications, etc. As the project moved from a development to a test phase, the need for accurate and organized feedback became increasingly important, and the role then took on qualities of mediating between users' needs and available programming resources. This same situation continued throughout the maintenance phase of the project.

The project manager also carried normal administrative responsibilities, such as scheduling, budget projections, plus the tasks of eliciting program and system documentation, as well as producing quarterly and final reports to sponsoring agencies.

3.3.3.2 Doctoral Intern

The doctoral intern staff has thus far borne a major responsibility for developing patterns of introductory and advanced academic usage of the Laboratory. In this role they worked very closely with all faculty associated with the project. The scope of the doctoral intern's role included serving as a teaching assistant in advanced seminars, acting as guest lecturer in colloquia and introductory courses, and planning for further integration of the Laboratory into the intermediate and advanced curricula of the School of Librarianship. Thus far, the graduate interns have taken on the major responsibilities for developing academic exercises and Users' Guides for the Laboratory, especially for the LABSEARCH system. For the intern, this Laboratory offers a useful opportunity to lecture, teach, and become involved in academic planning.

3.3.3.3 Laboratory Assistants

The third staff level required for the Laboratory's operation was the Laboratory assistant group. These assistants were Library School students who were especially interested in automated bibliographic methods or computer assisted instruction. The duties of the Laboratory assistants consisted of two main tasks. The first was to attend or be available during regularly scheduled Laboratory sections. The purpose of this was to have someone available to answer questions and help students to work with the terminals. The second Lab assistant task was to work with faculty and Laboratory staff on developing and testing academic exercises. This involved eliciting original material for exercises as well as pre-testing the properties of the specific programs that might be used for exercise material.

Training for lab assistants consisted of familiarization with the programs that the students would be using, including both error-correcting techniques and working out procedures for orderly communication with the Campus Computer Center when necessary. In the training of the assistants, we also stressed their educational and morale-supporting functions in terms of actively initiating discussion and encouraging

student comment and criticism. In this connection, we found one procedure to be especially helpful. Before a problem set was distributed to a class, we took the assistants through the problems, explaining the objective of the exercise, and the reasons why the problems were structured and sequenced as they were. At these briefings the assistants would sometimes criticize the exercises and make suggestions for their improvement. Thus, by the time they introduced the exercises to the students in their Lab sections, their familiarity with the problems went deeper than just "knowing how to get the right answers," and they were able to give more insightful and knowledgeable guidance.

3.4 Student Sessions

The first regular use of the Laboratory occurred in October 1969 with a group of six students enrolled in a research seminar in Advanced Methods for Intellectual Search and Access to Information. This first experience was, however, very limited both in the availability of Laboratory resources and in the size of the student group. The experience did, however, assist us materially in setting up protocols and procedures for later periods of more extensive use of a greater number of programs.

Beginning in January 1970, we opened up the Laboratory on a more extensive basis, for the Winter and Spring academic quarters. Three courses were the main users of the facility: two reference courses (elementary and intermediate) involving about 170 students, and one advanced research seminar involving about twelve students. About fifteen hours a week were reserved for organized Lab sections, with each student having a permanent assignment to a particular section throughout the quarter. Additional hours were also provided during peak periods for individual student use and for program development, system maintenance and debugging.

With three terminals available, we attempted to structure Laboratory sessions to include six students at a time, working in pairs at each terminal. Our decision to have the students work in pairs at the terminals was more from educational motives than from any administrative concern to get more usage out of available time. Most of the Laboratory programs were not conventional "computer assisted instructions" (CAI) systems, but were models of formal procedure. In such programs the student initiates the dialogue, decides on strategies and poses questions to which the system responds. This requires that the students provide a good deal of creative input while they are at the consoles. Our assumption was that students could get more out of their scheduled Laboratory time by working as a team, making shared decisions and dividing up the mechanics of typing and keeping a written record of the transactions.

In addition, we hoped that the partner arrangement would improve student morale. The impression that students receive from their initial contacts with the terminals can have a deep influence on the attitude toward technology which they carry with them throughout their professional careers - an important consideration for people entering a profession

which has an intensive ongoing concern with the utilization of technology. If the classes were to gain maximum benefits from the Laboratory, the time spent in it would have to be purposefully directed with as little distraction or interruption as possible. In particular we were anxious to avoid the kind of demoralization that often overtakes students in elementary computer sciences courses, when they find themselves spending long hours of drudgery at keypunch machines in the solution of what they believe to be relatively trivial exercises.

We also tried to facilitate matters for students by creating a controlled environment consisting of both exercises and users' guides. The exercises were designed to teach students specific material related to the program systems available in the Laboratory. Attention was focused on the step-by-step development of increasingly sophisticated procedures for exploiting program capabilities, with only minor emphasis given to the mechanical details of terminal technique.

We believe that using on-line programs to achieve educational goals can be seen as a three phrase paradigm:

a. Familiarization. What is the program like? How do you formulate acceptable input, what does the output look like, and how do you interpret it?

b. Exploration. How far can you push the system? What is the real range of its capabilities?

c. Logical command of strategy. Knowing the routines and the range of possibilities, what is the most effective way to make use of this knowledge for a variety of problems?

This paradigm forms the fundamental pedagogic core of the exercises developed for students working in the Laboratory.

Exercises can be considered as a way of gradually building a transition from uncertainty and unfamiliarity with automated and on-line procedures, to learning the material embodied in these procedures. Especially for Masters Degree level students, a series of connected questions in familiar exercise form can be the most effective way of spending a brief amount of time in the Laboratory. Frequently each exercise concluded with a "free play" question which invites the student to use the resources of the Laboratory to explore topics of his own interest.

The development of Laboratory exercises was a sophisticated and complex job, involving the efforts of both faculty and Laboratory staff. The exercises had to fit into a specific curriculum and lecture series, and each exercise was tailored to a specific course and instructor.

Laboratory exercises represent a broad-brush approach to the Laboratory, and embody only a small subset of all the possible combina-

tions of query, strategy, and retrieval. Further, these exercises, while serving as a useful introduction, do not necessarily equip the student to strike out in a self-directed way and explore the system according to his own interests. To remedy this problem we invested a considerable effort in the development of complete Users' Manuals for each of the four major Laboratory programs: MARC Search (CIMARON), Associative Search (LABSEARCH), Reference Search (REFSEARCH) and (DISCUS). (We have included these four Users' Manuals as independent volumes of this report.)

Our major goal was to write a set of manuals which would serve as independent reference guides for the student and staff users of the Laboratory. At a minimum, each Users' Manual was designed to provide operating instructions for students who actually are using the terminals, so that students need not initially memorize detailed program commands. Each of the Users' Manuals also attempted to include extensive discussion of the basic logic of the program system.

The most extensive of the Users' Manuals is the LABSEARCH volume. LABSEARCH is a program for associative searching, but associative searching is itself an advanced bibliographic notion for most library school students. The manual thus has to introduce the students to the bibliographic concepts which motivate the program as well as the operating details of using the program itself. The LABSEARCH program places a great deal of control in the hands of the user, and requires more active and informed judgments on the part of students. The text was therefore organized into a large number of short, clearly labelled sections, providing generous cross-referencing and indexing features, with numerous summarizing and overview passages, reference tables, and the like.

A similar difficulty is faced by the DISCUS Users' Manual. This is primarily a guide for analysts trying to write CAI courses. The introduction of computers as an instructional medium has created a new type of author: the scholar preparing a "textbook" with the express intention of having it used interactively by a student with access to a computer. Standard publications on technical writing do not address this problem, nor does the current literature on "how to write a CAI program." The primary aim of the DISCUS Users' Manual is thus to help the instructor* to prepare a sound, orderly, and attractive exposition of subject matter. The description of how the DISCUS programming language works is secondary to the larger conceptual issue of how to organize an exposition of subject matter in a way that takes advantage of interactive computer operations.

3.5 Academic Role: Integration with Curriculum and Research

The inventory of programs currently available in the Laboratory has been developed for both research and educational applications in

*Meredith, Joseph C. The CAI Author/Instructor, Inglewood Cliffs, New Jersey: Educational Technology Publications, 1971.

library and information science. In both cases, the Laboratory exists primarily as a resource for doctoral students and for faculty, and consequently there has been much effort towards integrating the Laboratory into existing research and educational programs.

With respect to teaching, we used the following procedure. Members of the Laboratory staff met with faculty members to explore ways in which specific courses might make use of the programs of the Laboratory. Decisions were made concerning the potential extent of Laboratory usage for each course, how the educational objectives of the course might be supplemented through Laboratory programs, and just exactly at what points in the presentation of course content Laboratory work might be introduced. The objectives of various groups of student users differed according to the nature of the course within which the Laboratory work was assigned.

Since formal lecture hours were scarce, we were anxious to get students onto the terminals as quickly as possible without extensive preparation or orientation. To achieve this end, we relied heavily on Users' Manuals and Laboratory assistants, and used only a single hour of lecture time to provide a general orientation to each student group.

This one-hour presentation was organized into two parts. In the first half-hour we described the characteristics of on-line interactive systems, explaining why such systems had potential for library service, and then introducing the Laboratory as an example of such a system, calling attention to the analog between an on-line system for educational purposes and one that might be implemented as a component of a library network. The second half-hour was devoted to describing the particular program that the class would be using in their Laboratory work. The program was discussed from a logical point of view, with stress on the formal aspects of its design that were related to practical questions of library searching. Little emphasis was given to the technical details of the program's internal organization. The presentation was geared to the logical rather than the engineering aspects of the programs, which in turn made it possible for students without scientific backgrounds to develop a preliminary understanding of the major design characteristics of the program in a short period of time.

It seemed best to us that instructors set the major objectives of the Laboratory experience and directly supervise not only the design of the exercises but also some of the actual Laboratory sessions in order to evaluate the technical and educational results directly. However, this approach requires that an instructor have the released time in which to plan and monitor sessions, and also preferably to participate in the evaluation, design and development of new Laboratory programs. Faculty advice is an invaluable support for the Laboratory, without which its existence cannot be maintained.

But faculty time is a scarce resource, and it may impose a burden on an academic department to detach an instructor from enough of his usual teaching, research, and administrative responsibilities to carry out these tasks. The inevitable compromise is, therefore, that some

programs are utilized by faculty members who had taken no direct part in their initial design and development, but who, upon seeing the programs demonstrated, saw uses for them which were rather different from what we had envisaged when we first set them up.

It therefore becomes necessary to think of Laboratory programs as routines which clarify and explicate various fundamental problems of bibliographic processing, and to keep their characteristics sufficiently flexible and general so that they may be adapted and incorporated into a variety of different courses taught by instructors who have varying educational motives for using the programs. This allows faculty participation to be broad-based and advisory, with the burden of generalization to be shared equally between users and Laboratory staff.

For doctoral study, the Laboratory serves as a tool for students' independent work on problems in bibliographic and information processing. In this capacity the Information Processing Laboratory is used as a system which allows the student to control search decisions, and permits direct comparisons of different decisions applied to the same data. The result is a powerful tool which enables students to organize and conduct an investigation of an automatic system. Our experience thus far strongly indicates that the on-line capability of the Laboratory enables students to develop a sense for the fruitful organization of such projects in considerably less time than otherwise would be the case.

3.6 Description of Terminal Hardware System

3.6.1 System Configuration

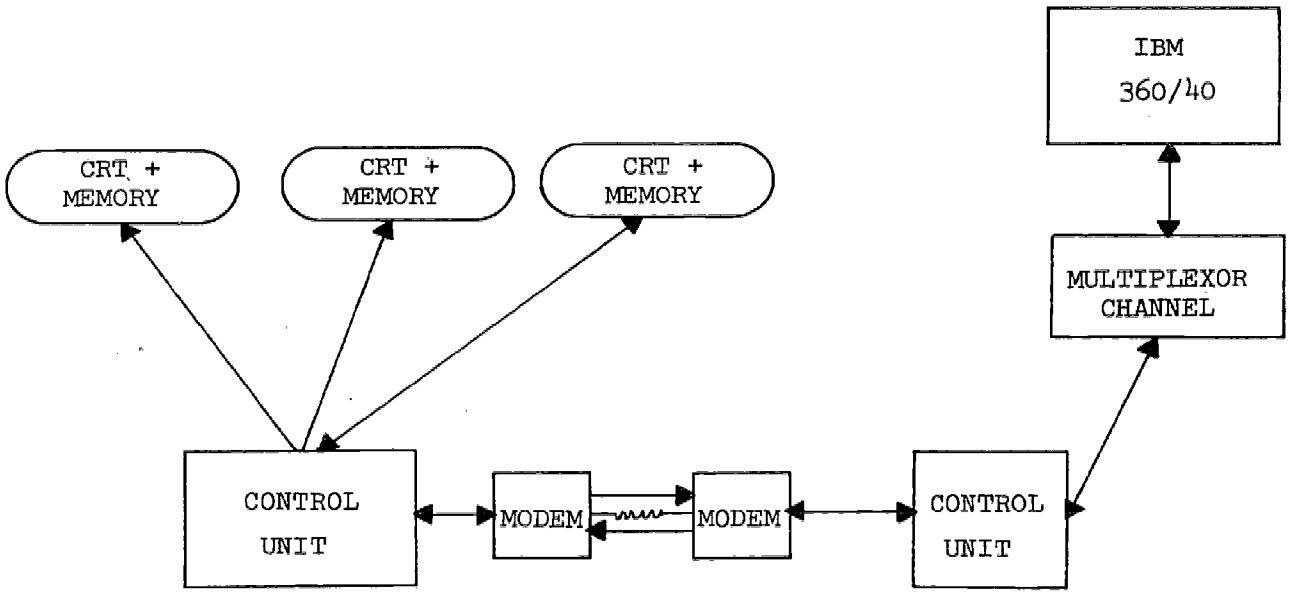
The basic hardware system of the Information Processing Laboratory consists of three video (CRT) terminals, plus a communications interface with an IBM 360/40 remotely located in the Berkeley Campus Computer Center. The three video terminals are part of a Sanders Model 720 Communications System. This system consists of a special communications interface, a memory, and video/keyboard terminals. Thus, the memory (and the screen) may be controlled by the user at a keyboard or by the computer sending messages across the communications link to the memory and then to the CRT screen. Communication between the user and the computer is effected by the user's decision to transmit the contents of the screen (i.e., the memory) to the computer. Transmission does not occur without this decision.

The Sanders 720 Communications System in the Laboratory includes a separate 1,024 character delay line memory for each video terminal. Every 21.5 milliseconds the entire memory is read out and used character-by-character to regenerate the display on the CRT screen. A CRT translation module of the 720 system positions the beam on the screen of the CRT as directed by the alphanumerics contained in memory. A constant flicker-free display of characters on the screen is maintained by regularly refreshing each screen character 46.5 times per second. The position of the beam and the characters is guided by the format control characters.

Thus, communication from keyboard to memory is by standard character input and screen formatting functions. Communication from memory to screen is accomplished as described above. Transmission from the 720 memory to the IBM 360/40 is initiated by the SEND BLOCK or SEND PAGE pushbuttons on the keyboard. There is a distinct 1024 character memory for each video terminal in the Sanders 720 system. A SEND BLOCK or SEND PAGE command causes the partial or total contents of a memory of a specific terminal to be transmitted to the computer. Similarly, when the computer sends data, each terminal (memory/screen) is addressed separately.

Although there are three terminals (separate memories, screens, and keyboards), the entire configuration is linked to the computer by two fixed circuit telephone lines. Control units within the 720 system handle polling, queuing and synchronization problems which arise with respect to handling three active Input/Output stations over the communications lines. Separate GE TDM-220 Data Sets are utilized at each end of the telephone lines to condition binary data for transmission and reception over the circuits. These units are known as MODEMs (modulator/demodulator), and are interfaced to two 4-wire Schedule 4 voice-grade lines which run about one cable-mile from the Laboratory to the Computer Center (actual distance - 400 yards). The transmission rate is 2400 baud or 300 characters per second. Thus three seconds are required to transmit an entire screen (1,024 characters). A Sanders control unit at the computer end of the transmission interfaces the message signal to the IBM 360/40 multiplexor channel, during both send and receive conditions. The following illustrates the flow of information.

FIG. 2: FLOW OF INFORMATION BETWEEN CRT TERMINAL AND IBM 360/40 COMPUTER



3.6.2 Terminal Display and Keyboard

Each terminal resembles a typewriter keyboard attached to a twelve-inch television (CRT) screen. Sixty-four different visible characters, consisting of upper case letters, numbers, and special symbols, may be generated on the screen area. In the Laboratory configuration, each screen may display characters (a maximum of 1024 [12.2 lines] may be displayed at any one time) within a matrix consisting of 84 character positions per line and 32 lines on the screen.

In addition to the 64 alphanumeric characters, the Sanders 720 system also has three major Format Control characters:

FORMAT CONTROL	FUNCTION
Vertical Tab (VT)	Position next character four lines down and flush left
Carriage Control (CR)	Position next character one line down and flush left
Horizontal Tab (HT)	Position next character four character positions to the right

Note that the normal 1024 screen characters may be distributed anywhere over the 2700 possible screen locations by using CR, HT, and VT control characters. In this way a limited number of screen characters may access and use the entire screen matrix.

The Sanders 720 keyboard resembles a conventional electric typewriter keyboard with the addition of an extra bank of pushbuttons on the right side. All the conventional keys are present, plus some special control keys. Data always appears in upper case, without using the shift key. When an alphanumeric key or spacebar is struck at the keyboard, the corresponding character or space is displayed immediately on the screen. In addition, a small blinking line appears at the bottom right of the character. The blinking line is referred to as the cursor. It performs the same function as the writing indicator on a conventional typewriter; that is, it indicates where the next typed character will appear on the screen.

Of the special control keys on the standard keyboard, the following are most useful to understand in formatting messages to the various Laboratory programs:

KEY SYMBOL	FUNCTION
SHIFT	When depressed, the entire keyboard is shifted into "upper case" mode.
REPEAT	When used in conjunction with any other alphanumeric key it causes that symbol to be generated continuously on the screen.

KEY SYMBOL	FUNCTION
SPACE	In lower case mode, this key duplicates the space bar; when the keyboard is in upper case mode, it serves to backspace the cursor. Both the space key and the space bar can be used in conjunction with the REPEAT key.
CR	CR moves the cursor up to the end of the previous line when the keyboard is in non-format mode. CR moves the cursor down to the beginning of the next line of data, when the keyboard is in FORMAT mode.
HOME	Moves the cursor back to the first alphanumeric character in the screen.

In addition to special control keys, the Sanders 720 system keyboard contains several Edit Function pushbuttons to the right of the keyboard. These pushbuttons establish the mode of operation of the terminal, and are in effect until another mode of operation is established. The data clear and data transmission functions are also located in this bank of pushbuttons. The Sanders 720 system employs three general classes of functions to allow origination and modification of data. These are: type, insert, delete. The type function allows replacement of data - i.e., data entered from the keyboard replaces previously stored data or spaces. The insert function allows addition of new data without destroying previously stored data, i.e., the previously stored data is spread to accept new data. The delete function allows selective removal of previously stored data - i.e., the selected data is deleted and the remaining data is closed up to eliminate the gap which results from the removal of the selected data. A fourth editing function provides for moving or resetting the cursor position. Normally editing is performed simply in conjunction with alphanumeric characters. The 720 system also has a parallel set of editing functions to be performed on format control characters as well as alphanumeric characters. To accomplish this, the keyboard must be in FORMAT mode.

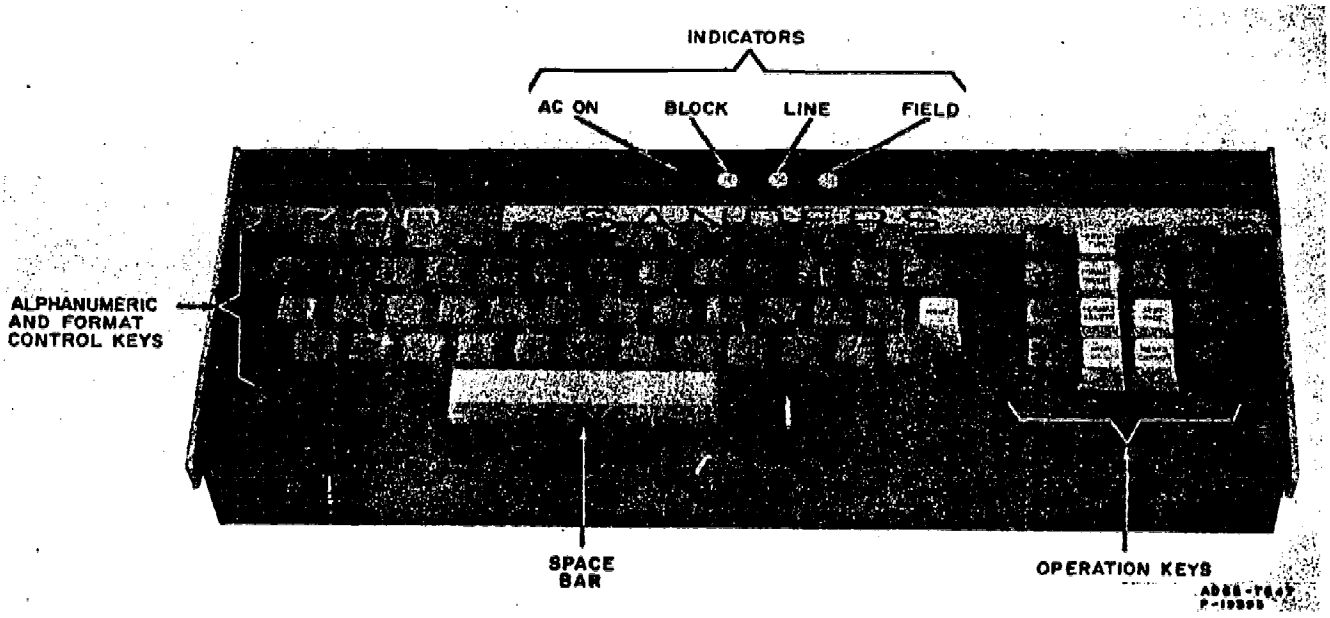


FIG. 3:
SANDERS TERMINAL KEYBOARD

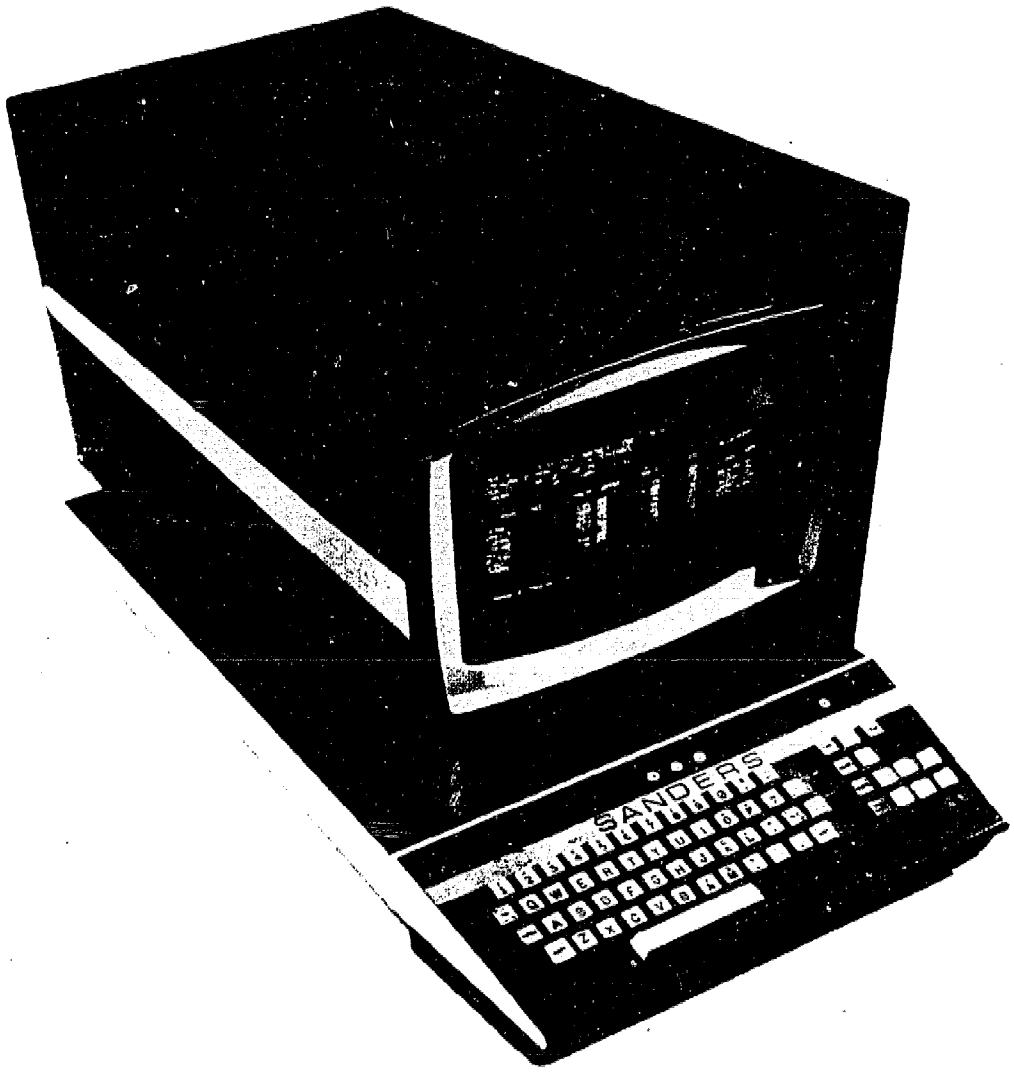


FIG. 4:
SANDERS MODEL 720 VIDEO DISPLAY TERMINAL USED IN INFORMATION PROCESSING LABORATORY

4. THE INFORMATION PROCESSING LABORATORY AS AN EDUCATIONAL RESOURCE

4.1 Introduction and Summary

In the previous chapter we discussed the gross structure of the organization and environment of the Information Processing Laboratory. In this chapter we describe in some detail our practice in using the Laboratory as a resource for teaching.

The content of what is taught in the Information Processing Laboratory often deals with controversial topics such as the potential role of the computer in libraries and in information science. However, the basic pedagogic premise is still quite conservative and in consonance with other library school courses. Namely, that wherever possible, learning should be realized in the direct experience of performing exercises, solving problems, and making observation under controlled conditions. The Information Processing Laboratory thus provides a facility for the information science curriculum which is analogous to the service that the cataloging and reference laboratories provide for their curricular counterparts. Students are given the opportunity to work independently, observing the application of conceptual principles which have been presented to them in lectures and textbooks, and the introduction of exercises and laboratory materials is carefully coordinated with the requirements of the academic program.

In the sections that follow, we will discuss four of our major laboratory systems from an educational point of view. Two of these systems, LABSEARCH and CIMARON, are typically used for information science oriented courses, while the other two, REFSEARCH and Machine Tutorial Mode (MTM), were created in response to the faculty's interest in developing an approach to traditional library practices that would put greater stress on formal interpretations of commonly encountered professional procedures.

Two of the major Laboratory program systems are oriented to the information science and library automation aspects of the School of Librarianship curriculum. These programs are the Associative Search System (LABSEARCH) and the MARC File Search System (CIMARON). Courses in these areas are quite dependent upon the ability of students to carry out exercises and projects which require the direct experience of the operational computer systems in the Information Processing Laboratory. For most students this is an opportunity which is unique in their educational careers, and is likely to remain unique even through a great deal of their professional careers as well. The descriptions of LABSEARCH and CIMARON in this chapter will emphasize the role of these programs in an educational context: what topics are they used to teaching, how Laboratory assignments can be used to deepen insight and extend understanding.

The potential contribution of the Information Processing Laboratory extends beyond the direct limits of information science in librarianship. To explore this potential, we experimented with Laboratory systems which could be of service to areas of traditional library education, such as reference and cataloging. We attempted first to reorganize along more formal lines. The presentation of the traditional systems has led us to try to extend this methodology of reference-question and reference-work curriculum.

The REFSEARCH system, which was the embodiment of our experimental work in this area, represented an attempt at formalization more than an attempt to automate reference services. As an educational device, REFSEARCH provides a valuable adjunct to the traditional syllabus-oriented method of teaching general reference.

The Information Processing Laboratory Project also experimented with using computer technology as an innovative method in education. For this work we created our own version of Computer Assisted Instruction (CAI), which is more suited to the tutorial methods common in humanities courses. Within this Machine Tutorial Mode (MTM) we selected Subject Cataloging as an initial area for which to develop machine-based courses which could be "taken" by students independent of extensive faculty contact. During this current phase of the Laboratory project, we developed an efficient language and compiler for writing terminal-oriented MTM material. We also converted and amplified our Phase I Subject Cataloging CAI course to the new video terminal-oriented compiler. These two topics are extensively discussed in section 5 of this chapter. A monograph describing the Laboratory CAI system has recently been published, and interested readers are referred to that source for additional information.*

In summary, the basic goal of the Information Processing Laboratory has been service to teachers and to students. A primary area of interest was information science and library automation, although this did not exclude the development of systems to further education in traditional library subjects. In this chapter we will discuss the educational relevance of four major Laboratory systems (Associative Search, MARC File Search, Reference Search, Computer Assisted Instruction). Detailed operating instructions for these four systems are to be found in four separate volumes of this Phase II Laboratory report.

4.2 Associative Search System (LABSEARCH)

LABSEARCH is the name given to a family of interactive information retrieval programs which currently constitute a major segment of the Information Processing Laboratory system. These programs are based primarily on the principles of search and measures of index association as elaborated by J.L. Kuhns.** The LABSEARCH system represents sophisticated techniques of document searching and of processing retrieval requests. The system is used in advanced seminars in the School of Librarianship as well as for independent research.

*Meredith, Joseph C. The CAI Author/Instructor, Inglewood Cliffs, New Jersey: Educational Technology Publications, 1971.
**Kuhns, J.L. "Continuum of Coefficients of Association," Statistical Association Methods for Mechanized Documentation, National Bureau of Standards Miscellaneous Pub. No. 269.

In this section we will give a brief description of the mechanics of the LABSEARCH system, emphasizing the role of LABSEARCH in the context of School of Librarianship seminar 242AB (Formal Methods of Intellectual Access to Information). (A separate volume in this current report entitled LABSEARCH User's Guide, by Edmond Mignon and Irene Travis, contains a detailed explication of the commands and options available to LABSEARCH terminal users.)

4.2.1 Program Description

The most distinguishing feature of the LABSEARCH system is the interaction permitted between system and user. Search requests can be formulated in terms of subject descriptors connected by Boolean operators including parenthetic nesting. Selected elements of a search request can be emphasized by means of assigning weights to individual subject descriptors or to groups of descriptors. The system user controls the extent of searching by specifying the type of association coefficient to be used during the search procedure. Search results can be ordered in ascending or decending order of probability of relevance, and searches can be repeated by simply varying the type of association coefficient to be used. The major portion of LABSEARCH's interactive flexibility is implemented by means of a special command language.

4.2.1.1 Data Base

The current LABSEARCH data base consists of approximately 400 records, representing a professional journal article in the field of library and information science. The data base was selected from materials published during 1957-1965. Each records consists of:

Microfiche Copy of the entire article or report. Students are encouraged to make microfiche copies for their own collections.

Abstract of the article or essay. When available, author or review journal abstracts are used. Otherwise, abstracts were written by Laboratory staff and students. The abstracts are keypunched and stored in machine-readable form, and can be displayed on the video terminals.

Bibliographic citation data (author, title, source, date, etc.) are also machine-stored. The author-title portion of the bibliographic data can also be displayed on the Laboratory video terminals.

Subject Descriptors are drawn from a Subject Authority List of 350 terms (see LABSEARCH User's Manual) and are assigned by members of the Laboratory staff. The documents are indexed with an average of 15 terms per document. The subject descriptors for each record in the collection are machine-stored and form the nucleus for computing measures of association and for processing retrieval requests.

Association Files. The LABSEARCH retrieval logic is based on coefficients of association which are computed from the statistics of index term assignments. Different files of association coefficients

are computed by off-line batch-operated programs, and the resulting association files are made available to the interactive search routines of the LABSEARCH system. An association file consists of a head-term (each of the 350 terms of the Subject Authority List) plus four other index terms which are most highly associated with the head-term. For example, in the current collection, according to the coefficient of association known as KUHNSW, the following association file entry exists for the term GRAMMAR.

Term	Association Coefficient
GRAMMAR	(Head Term)
PARSE	.45
SYNTACTIC ANALYSIS	.44
FACT RETRIEVAL	.41
SET THEORY	.41

Altering the contents of the collection or changing the method of computing the coefficient of association results in a different entry for GRAMMAR. Currently eight different association files exist, each corresponding to a different method of measuring the relatedness of documents as described by Kuhns (op. cit.)

4.2.1.2 Major User Commands

LABSEARCH asks six questions during a normal pass through the program so that both the association file and the search request can be entered by the user. These questions are:

Q01 Do you want word association?

Q02 Specify association file.

Q03 Do you want scoring?

Q04 Enter Boolean expression.

Q05 Do you want results displayed?

Q06 Specify restart or exit.

The questions are self-explanatory and the answers given by the user are straightforward. However, at any time the user can switch to a command language in order to exercise other program options. The command language consists of the following major functions:

DISPLAY: used to display on the terminal screen either association files; e.g. (DISPLAY 'GRAMMAR') or document accession numbers which are the results of a search request (DISPLAY DOCUMENTS).

GET: used to examine the index descriptors of records in the collection (GET 'A0121').

GO TO: used to branch to one of the normal pass question points (GO TO Q04).

RETRIEVE: use to display the abstracts of records in the collection (RETRIEVE 'A0121') or of records in a retrieval request (RETRIEVE DOCUMENTS).

SORTA: used to sort retrieved records by relevance scores in ascending ("lowest first") order.

SORTD: used to sort retrieved records by relevance scores in descending ("highest first") order.

4.2.2 Educational Relevance

It is difficult to overstate the fundamental importance of searching to professional library practice. We see the ability to search effectively as the key skill that distinguishes the expert user of bibliographic files from the inexperienced one. This skill is a notable competency of veteran librarians and scholars with many years of experience, and it has been claimed, not altogether unreasonably, that there is no fully adequate substitute for the bibliographic wisdom that comes from many years of practice. But we believe that the interactive character of the Information Processing Laboratory provided us with a means for accelerating the cultivation of this skill by giving students the opportunity to conduct searches far more rapidly and conveniently than with manual systems. In the Laboratory setting the student specified the search procedure that he wished to carry out, but the actual execution of the search, which is a lengthy affair in a manual system, was rapidly done by the computer. The focus of the student's attention and the major part of searching time were thus shifted from the mechanics of performing the search to the logical and intellectual questions of choosing an effective procedure and interpreting the consequences of such choices.

We chose to include and emphasize statistical association in our most flexible retrieval program, LABSEARCH, partially because it lent itself to automatic methods, and more importantly because it was a new and provocative approach to searching which could not be presented effectively without the Laboratory, and was thus an ideal example of the use of the computer for augmenting and modernizing the professional curriculum. However, there seemed little point in training classes to conduct associative searches just for the sake of exposing them to a retrieval procedure whose suitability for any practical service situation was undetermined. Our goal therefore was to use associative searching to introduce the overall

logical and interpretative problems of search organization by the direct experience of observing different search strategies. The educational benefits extended, therefore, to the general objective of developing an understanding of the fundamental qualities of search and access to information.

4.2.3 Educational Procedure

Information access and searching were presented in connection with class lectures on the association measures of J.L. Kuhns as a new tool for document search and retrieval. In the Laboratory sessions, we were concerned with the double objective of giving the students the opportunity to witness the behavior of the association measures and also to develop a model of search procedure.

The second of these objectives presented a methodological problem. An effective search requires many acts of judgment. Choices from a wide range of strategies are made and introduced at various decision points, relevance judgments must be made, rankings established and revised, and so on. We do not rule out the possibility that with further experience we may discover that the principles by which such judgments ought to be made are straightforward and easily explained. But if such principles exist, they are not yet clear to us, and thus the organization of an effective search so far remains an apparently subtle, complex, and slightly obscure matter.

We therefore decided to start the classes with some exercises that would be conceptually more elementary than even a very simple search; they would be preparation for searching. This may seem strange, since our classes were made up of advanced students who already had some experience with searching from their previous studies. But the searches in this class were to have a different emphasis, since here we were not interested in requiring the students to produce acceptable outputs for specific information seeking problems, but rather to develop a better understanding of the searching process itself.

We adopted a convention for the elementary exercises which was of considerable pedagogic convenience: we specified in advance the documents that were to be assumed as relevant to the request. With this device, we were able to defer until a later stage the difficult questions of how to determine relevance, or of how to know when to quit. This convention enabled us to reverse the "real" situation and provide a more tractable setting for learning. The "real" search started with a query and ended with some selection of suitable documents; the exercises started with a specified set of suitable documents, and worked backwards to show the variety of strategies that may lead to the retrieval of all or some of these items. In this way we could focus on various aspects of procedure without introducing the complications of requiring the student to produce a "correct" output, or of asking a class to debate and come to an agreement on what such an output should be.

In addition to providing the students with the "relevant" set, we also broke up the total problem, highlighting individual choice points

and techniques. These were introduced one at a time, and discussed in some detail before the students were asked to put all of these techniques together into an organized procedure. For this purpose, we chose an interpretation of searching which did not represent a fully developed formal account of the process, but which served as a convenient way to segment it for teaching and learning purposes.

Typically, document collections are likely to contain items which are wholly or partially relevant to a query, but which are not retrieved due to a lack of match between their index terms and the specifications of the request. It is, therefore, of interest to broaden the range of the query in the hope of locating additional relevant documents. This is often thought of as a process of compensating for misjudgments in indexing. In the class, however, we took the position that the expansion of a request, whether by statistical association or more conventional means, serves the purpose of placing the items retrieved by direct match into perspective. This step is desirable even if the indexing is altogether adequate. Imagine, for example, the extreme case where in fact there is really nothing in the collection even partially relevant to the request that is not retrieved by direct match. Even so, it will still be necessary to expand the request in order to confirm that this is truly the case.

Associative retrieval gives us a formal and automatic method for providing this perspective for the direct match documents, thus providing the system's estimate of the area of document space that will provide the most useful framework of relevant, partially relevant and nonrelevant documents for helping the searcher to determine how (or if) the search is to proceed. In a previous study* we pointed out that the more extensive retrieval brought forth by searching in associative mode results in a loss of precision. A year later, it now seems to us, that far from being a criticism of associative searching, this is just as it should be, at least for educational and research purposes, because the partially relevant and nonrelevant items are needed in order to give the searcher a vantage point.

Once the request has been expanded in this way, the student can proceed to narrow the output once again, by determining the degree to which the retrieved items satisfy the query, identifying the properties that distinguish the satisfactory items from the unsatisfactory ones, and going on to examine the extent to which these properties are reflected (or can be estimated) from the indexing. At this point, various techniques for contracting the output, such as thresholds and weights, are suitably introduced. In a "real" search, a user would not be likely to go back and resubmit an improved variant of his request once he had identified the useful items; he would simply note the references that satisfied him and be on his way. But when the concern is with understanding

*Maron, M.E., A.J. Humphrey, and J.C. Meredith. An Information Processing Laboratory for Education and Research in Library Science: Phase I, Berkeley: Institute of Library Research, University of California, July 1969.

the search process itself, it is of far less importance for students to get a satisfactory output than it is to account for their results.

To summarize, we saw three check points in the search as being of particular interest:

a. Request formulation

b. Request expansion

c. Refinement of the expanded request

We paid less attention to the first of these considerations, because it was less dependent on the on-line methods for effective exposition, and the students already had some understanding of the problems associated with this step from previous course work in bibliographic organization. But the examination of the second and third of these topics would be new material, and our hope was that after some study of these steps, the students would be on their way to a more careful understanding of the search process, and their experience in the lab would give them a more detailed model of the procedure than could have been obtained from classroom experience alone.

4.2.4 Analysis of an Introductory Exercise ("Associative Searching Warm-up Exercise")

The exercise described below is introductory.* It is intentionally a bit casual and unchallenging, but nevertheless it covers a good deal of material. Direct and associative outputs are compared, ranked output is called for, and association tables are summoned and compared. Thus all the basic manipulations are introduced. On another level, the exercise helps the student to get used to the kind of index term relationships that are typical of the association tables. This involves a minor adjustment of the student's habits before he can interpret the tables, and perhaps calls for a word of explanation.

From their experience in searching conventional files, students are accustomed to think of request expansion in terms of considering synonyms or hierarchical term relationships. They are used to displays like

Grammar, see also

Dialect

Morphology

Syntax

from which one chooses the alternative term which seems most suitable for a particular need. But in statistically associated term relationships, the closest term pairs are computed without regard for the meaning

*The text of this exercise is given in Section 4.2.7.

of the words. In practice, however, the term pairs that result from this procedure are often intuitively reasonable. Their connection is not one of synonymy, but rather one of likely co-occurrence, e.g.

Grammar (is statistically associated with)

Parse

Syntactic Analysis

Fact Retrieval

Set Theory

Parse and Syntactic Analysis are not synonyms for Grammar or for each other, but they serve the useful purpose of reporting to the searcher the indexing environments in which the term Grammar has been most distinctively applied. The environments offer the searcher clues about the direction in which to move in the index space in order to retrieve more documents of interest. This notion of term relationships is not altogether new--we make use of it in consulting a conventional thesaurus of the Roget type--but there is some novelty in the idea of systematically exploiting it for bibliographic searching. Even when synonyms are in fact present in the Subject Authority List, they may not co-occur in the index term sets of the documents, and then may have a low correlation. The key adjustment that the student has to make, then, is getting used to the idea that in an association table he will not as a rule find synonyms for the header term.

In the first two steps of the exercise* the student submits the same request in both direct match and associative mode. Direct match retrieves two documents, and the associative expansion produces ten additional cites. Note that the request is simple, involving only one kind of operator, requiring no parenthesized expressions, and representing a plausible combination of terms. The third step of the exercise arbitrarily establishes a set of five relevant documents, following the convention discussed earlier. The student is then asked to make successive modifications of the original query, substituting increasingly less strongly associated terms for the original request terms, and discovering how far away he can move from the original request and still be able to retrieve the original set of relevant documents. By step 8, the substitute terms have a rather weak average association with the original terms (about .22), yet when association is called for all five of the relevant documents are retrieved, although only one of the relevant documents was retrieved by a direct match search using that request (Step 7).

Step 6 which asks the student to record in a table the two terms most highly associated with the request terms isintroduced to show that stongly associated terms frequently have plausible relationships. This point is not stated explicitly in the written text of the exercise but was discussed with the class when the exercise was being worked. The associated term pairings are: user/needs, information/communication, scientific/network,

*See Section 4.2.7. All subsequent mentions of "the exercise" will refer to the exercise contained in this figure.

and research/technology. None are synonyms, but all are reasonable clues to the context of the index terms.

In steps 10 to 12 other association files are introduced for comparison purposes. Filling out the table in step 11 reveals that the association tables produced by the two measures considered, KUHNSY* and DOYLE,* have almost no terms in common, thus providing a dramatic and somewhat extreme illustration of the considerable differences in the index term rankings produced by the different measures even though the raw data from which the tables are computed are the same for all measures. Remarkably, the effect on retrieval when KUHNSY is substituted for the DOYLE measure in step 10 is not as extreme as this situation might lead you to expect: the Y measure produces 4 out of the 5 relevant documents.

It is important to bear in mind that this exercise, although introductory, was devised for a rather advanced class which was approaching the topic of searching from the point of view of formal methods. The instructor had already discussed and compared a variety of association measures in lecture before the class began this exercise. This situation thus constitutes an exception to our feeling that in general it is best to ask a class to stay with a single association measure throughout the period when they are initially familiarizing themselves with the range of commands and routines available for on-line searching. Our experience has been that few students can organize a procedure for comparing two association measures until they have developed some sense for how a single one may be purposefully used. Or to put it more generally, skill in searching depends less on understanding the computational properties of a measure than it does on having a capacity to judge when and how to narrow or broaden a search. Until some facility is developed with this technique for readily interpretable and relatively simple cases, the student simply does not have enough substance to his experience to give him a useful frame of reference for making a comparative trial. Measure A may indeed be demonstrably better than measure B, and it may not be at all difficult to see that this is so, but what is educationally important is not to know this fact, but to develop some idea about how one might go about accounting for such a state of affairs.

4.2.5 Analysis of Exercise to Explore Precision Measurement

This second exercise, entitled Precision Devices,** focuses on the third of the three checkpoints that we mentioned earlier. The exercise takes about three terminal hours to complete, and is admittedly a bit lengthy

*For explanation of the KUHNSY and DOYLE measures of association see Mignon, E. and I. Travis. LABSEARCH: ILR Associative Search System Terminal Users' Manual, Chapter 5.

**For text of exercise see Section 4.2.8.

for its purpose, even though students were assured that there was no pressure on them to complete it by any deadline.

The first ten steps provide an introduction to the technique of narrowing the retrieval environment by reducing the number of associated terms. The exercise begins in a straightforward follow-the-instructions way, but then attempts to move the student as soon as possible into a state where he can make a few elementary predictions and interpretations, since such judgments will ultimately form the heart of his procedure when he advances to the stage of conducting and evaluating a complete search.

In the course of this exercise (steps 11-15) students are introduced to the use of term weights. The question of how weights should be assigned is rather elusive, and is usually baffling to beginners who find difficulty in finding a rational way to choose a weight of one magnitude in preference to another. In this exercise we offered one plausible interpretation of weights, although in lectures it was stressed that numerous other approaches might be just as good or better. The weight of .50 used in step 11 has very little effect on the ranking in this case but in step 13 the preferred documents are moved to considerably higher positions in the ordering by substituting a weight of .25. Multiplying the association values in the table for "information" by .5 does not reduce the scoring of the affected documents enough to produce a striking change in the ordering of the output list, but multiplying by .25 does lower the values of the documents containing that term enough to cause them to be ranked below the documents identified as relevant.

Steps 16-18 develop the idea of elaborating a request by using the NOT operator to cancel out noisy terms in the association tables. This is particularly useful when one of the request terms is rather general in meaning and has more than one context suggested by its associated terms. For example, the terms most highly associated with "match" are "question," "profile," "selective dissemenation," and "answer." Two contexts are suggested here, "question" and "answer" from the literature dealing with methods for matching input requests with stored data, and "profile" and "selective dissemenation" from the literature of user needs. To a person searching for documents in the latter of these interest areas, the high association of two terms from the former field with his request term will tend to lower the precision of his retrieval. However, the request can be modified by rephrasing it as

...'MATCH' AND NOT ('QUESTION' OR 'ANSWER')

and the precision will increase without sacrificing the extended retrieval capability of the associative mode.

What does an exercise like this accomplish? Primarily, it provides the student with a worked-out example of how common techniques may be applied, interpreted and related. But more generally, it is building up a set of experiences that will be shared by all the members of the class to provide a common basis for discussion and reference, and where the experiences have been staged so as to emphasize the properties of various tactics rather than the issue of getting correct answers or efficient

system performance.

The simple provision of a routine, for example, in which applying a weight to a request term produces an interesting and rewarding result, is welcomed by the inexperienced student, who, if left alone without any models and told to find his own example of a successful weighting, may spend many frustrating hours in the lab, and when it's all over still not have any results that he can discuss clearly with other students unless they take the time to repeat his particular sequence of trials. But if given an example of a successful application, he is more likely to have some belief in the power of the resource and will have more patience to eventually work through his own applications.

The problems chosen for the exercises do not necessarily represent the most clear cut illustrations of the points they are meant to elucidate. At the time they were chosen, they represented the clearest examples that we had discovered so far, but since our own knowledge of these matters was still developing, we were not in a position to determine what the "best" illustrations would be like. In presenting the exercise to the students, we stressed that the whole question of developing generalization about how far one can generalize about the value of a strategy that appears to be useful for particular situations is in itself a major object of investigation in the study of searching, and several discussion hours were devoted to consideration of procedure for specifying and testing such generalizations.

In this exercise, as in the one discussed previously, we were still using the device of asking the students to look at selected details of the procedure, without concerning ourselves with the total structure of the search, or the overall goodness of the results. The intent is to keep as many of the factors constant as we can, so that the student was not distracted by too many complexities arising from a variety of causes. The request itself is a bit contrived, but the interest lies in clarifying the consequences of particular classes of manipulations, (an approach for which interactive on-line situations are supremely well suited). The more advanced question of how such manipulations should be employed effectively is held in reserve until the class accumulates more experience.

4.2.6 Evaluation

Even though we have had several classes use the associative searching program, what we are reporting on is essentially a first trial, representing an intermediate stage in the development of our knowledge. We will need to use the program with more classes and instructors before we will have definitive views on how searching can be taught most effectively with the on-line capabilities of the Laboratory. The judgments which follow, however, even if they represent interim findings gained from our direct experience and characterize our collective view of the Laboratory project as a resource in advance education.

We cannot stress too strongly that in preparing the exercises, we ourselves learning about the system and also about the topic that we were endeavoring to teach. A major lesson has been the realization that skill

in presentation of the material comes only from continuous experience, and that such experience should be tied to direct faculty involvement. Our procedure in the 242 course was to explain to the students that our knowledge of how the material should be presented was still developing, and to encourage them to experiment with the terminals and report any results that they would recommend for pedagogic utility. We also organized our own concurrent effort, in which Laboratory staff were assigned to prepare exercises, to be reviewed and edited for presentation to the class. In the first of these procedures, the emphasis was on encouraging students to discover principles of searching, whereas in the second one the concern was to isolate specific situations that produced readily interpretable results suitable for teaching.

There were limits to both these efforts. The results of uncontrolled student explorations with the programs were almost uniformly disappointing. We believe that this is because it is not possible for most students to organize experimental trials such that results can be accounted for systematically and provide cumulative understanding. Frequently students attempted to evaluate outputs by varying measures and request conditions without being able to synthesize or interpret the results in a systematic way.

We learned from this that the very ease with which an interactive system can be operated means that it is easy for students to pursue many unrelated options without ever arriving at an educationally meaningful result. This freedom to explore is desirable, of course, because it encourages a freewheeling approach, and when used skillfully will greatly accelerate the rate of discovery and learning. But flexibility, convenience, and fast turnaround do not relieve the student or the instructor of the need for careful planning and for lucid models of organized on-line investigations. We now feel that far more stress should be placed on simple investigative techniques before students can be expected to use their unstructured terminal time purposefully.

In our second approach, that of the Laboratory staff designing exercises, we underestimated the time and complexity of the task. The design of good exercises and the sequencing of exercise objectives is not only a technical but also an academic problem, and a high order of expertise and experience is needed. The resulting exercises were not inadequate, but the time required to design and evalueate them should be taken into account. We made the mistake of thinking that once the program was operational, the majority of our problems were over. The effective pedagogic use of the program turned out to be an equally demanding job, requiring much clerical labor as well as advanced conceptual insight and complete mastery of the subject matter. It is, in our opinion, the natural context for increasing faculty role in the utilization of the Laboratory.

4.2.7 Associative Searching Warm-Up Exercises

Notation: Let T_1, T_2,...be the terms of the original request. Let T_{j1}, T_{j2},...be the terms most highly associated with term T_j, such that their association values A_{j1}, A_{j2},...have the relationship $A_{j1} \geq A_{j2} \geq \ldots$ Thus T_{j1} is the same as I'_j in the Maron-Kuhns notation.*

1. Submit a request in direct match mode for documents indexed under all four of the following terms: user, information, scientific, and research. Thus this request will be of the form:

$$T_1 \cdot T_2 \cdot T_3 \cdot T_4$$

Which documents satisfy this request?

2. Resubmit the request in associative mode, using the DOYLE measure and the scoring option. How many additional documents do you retrieve? List the document numbers in rank order. (Not necessary to list relevance values.)

3. Note the documents which have a scoring greater than .18. For purposes of illustration, let us arbitrarily consider these documents to be the items that are in fact relevant to the request. List the document numbers, with their relevance values.

Remark: The purpose of the next 6 steps is to illustrate the power of associative searching to retrieve relevant documents by means of the closeness of index terms. We are going to make modifications in the original request, and see how far we can move in the index space, and still retrieve all or some of the relevant documents (as defined in step 3).

4. Using the DISPLAY command, find the 4 terms most highly associated with user. Call these terms T_{11},...T_{14}. What is the association value of T_{11}? _____. Now substitute T_{11} for user, and repeat step 2. How many relevant documents do you retrieve? Which revelant documents do you fail to retrieve?

5. Repeat step 4, substituting T_{14} for T_{11}. Report your results.

*See Maron, M.E., and J.L. Kuhns, "On Relevance, Probabilistic Indexing and Information Retrieval," JACM, (VII, July 1960).

6. Using the DISPLAY command again, find out which term is most highly associated with each of the other three terms of your original request, and note what you observe:

T_2: information. T_{21}:_____ A_{21}:_____

T_3: scientific. T_{31}:_____ A_{31}:_____

T_4: research. T_{41}:_____ A_{41}:_____

7. Now submit this request in <u>direct match mode</u>:

$$T_{11} \cdot T_{21} \cdot T_{31} \cdot T_{41}$$

Compare your output with the result of step 3.

8. Same as step 7, only now using the DOYLE association file. Report your results in the same way as called for in step 4, but also compare the ranking of the output with that of step 2.

9. Make up your own variant of the orginal request and report your results. (Use a separate page if there is not enough room on this form for your report.)

10. Return now to the original request ($T_1 \cdot T_2 \cdot T_3 \cdot T_4$), and repeat step 2, this time using the KUHNSY association file. How do the results compare, with respect to a) ranking, b) precision, c) recall?

11. Using the DISPLAY command, find the terms that KUHNSY associates most highly with the request terms. You will have a total of at most 16 terms that are highly associated with your request terms. How many of these terms were also highly associated in the DOYLE file? _____. In particular, note the following comparison:

	KUHNSY		DOYLE	
j	T_{j1}	A_{j1}	T_{j1}	A_{j1}
1				
2				
3				
4				

4.2.8 Precision Devices Exercise

Three of the common methods for improving precision are:

1. Reducing the number of associated terms to be used in a search.

2. Weighting request terms.

3. Increasing specificity of the request by adding more terms with either the AND or the NOT operator.

The principles by which one of these procedures might be chosen in preference to the others are not yet well understood, but the problems that follow are intended to give you some idea of the effect of each of these devices on your output.

1. Submit the request

   'SCIENTIFIC' AND ('INFORMATION' OR 'LITERATURE')

   using the KUHNSL association file.

2. Retrieve the association tables for your three request terms and copy the tables down on a separate sheet of paper where you can have the information handy for each reference. (All three tables can be retrieved with a single input. Consult your manual if you don't remember how to do this.)

3. Using the SORTA command, [illegible] documents listed with the ones with the lowest scoring appearing at the head of the list. In columns 1 and 2 of the WORK TABLE, list the documents with the lowest relevance numbers, with their scoring.

4. Using the GET command, call for a display of the indexing of these documents. By comparing this indexing with the tables that you copied in answer to Question 2, note in columns 4 and 5 or 6 the index terms that caused each document to be retrieved.

5. Using the RETRIEVE command, make an estimate of the relevance of each of these eight documents to the request by reading their abstracts. Do not worry about the fallibility of your relevance judgments - that is not important for the point of this exercise. (If uncertain of the relevance of a document, consider it to be nonrelevant). Note your judgments by marking R or $\bar{R}$ in column 3 of the TABLE.

6. Now calculate a precision figure just for this subset of 8 documents. Remember that since we used the SORTA command, these documents represent the system's estimate of the documents least likely to be relevant that are nonetheless still worth retrieving. Hence you should not be discouraged if your precision is low.

   Precision = _____

7. Each association table lists the four most highly associated terms with each request term. Ask the system to research your request using only the three most highly associated terms, and now repeat the procedures of steps 3-6, using the lower part of the WORK TABLE to record your results. Some of this information will not be new, so you'll be able to copy some of your data from previous steps.

   Precision = _____

8. Compare your results with your first pass. Which documents have you lost? Why?

9. Using this information, can you now predict which documents would be lost if your resubmitted your request, asking for a search on only two most highly associated terms? Try it and see how good a prophet you are. (Not necessary to fill out the table for this step.)

10. Again look at the abstracts for the eight documents with the lowest scoring, and calculate the precision for this subset.

    Precision = _____

11. Refer to your association tables again. The terms highly associated with "literature," ("journal," "bibliography") are a bit closer semantically than the terms associated with "information," ("needs," "user") for the needs of a person who was interested in scientific information or scientific literature. This suggests that if "information" were assigned a weight, the output might give higher ranking to a greater proportion of relevant documents. Resubmit your request, assigning a weight of .5 to "information." Once again, use the SORTA command for the ordering of the output.

12. Referring back to your data from questions 3 and 4, which of the set of 8 documents were retrieved because they had a term associated with "literature"? How have their rankings changed?

13. Repeat steps 11 and 12, this time with a weight of .25 for "information." Report your results.

14. Why do you think the weight of .25 produces such markedly different results from the weight of .5? Do you think that a weight of .4 would have produced results more like step 12 or step 13? (Hint: consult the association tables.)

15. Does this suggest to you a rule that might help a user to determine what value to choose in assigning a weight to a request term? What would your rule be?

16. Returning to your WORK TABLE, which of the associated index terms is most frequently found to be assigned to documents that you felt were non-relevant?

Precision Devices Exercise (cont.)

17. Resubmit your request one more time, without weights, but adding the term you chose in step 16 to the request, by connecting to your original request with AND NOT.

18. Compare the precision this gives you for the 8 documents with the lowest scoring with the figures you obtained from previous procedures. Do you see any particular advantages or disadvantages to this method of improving precision?

You have now been introduced to three techniques for improving precision, and on the basis of a single trial you perhaps have some preliminary preferences for one method over the others. Choose one of the following two trials to validate or modify your first impressions.

A. The KUHNSS measure has a reputation for poor precision. Repeat the exercise, substituting this measure for KUHNSL, and see which of the techniques produces the most impressive improvements. Once again, for simplicity and conservation of effort, limit your evaluation to the 8 documents in each output with the lowest scorings.

B. Sticking with KUHNSL, repeat the exercise with another question. You are welcome to make up your own question, but KEEP IT SIMPLE so that you can interpret your results easily. If you prefer not to use your own question, here are some suggested searches which work well - but be alert for the possibility that they will not behave in every detail the way the question used in your original problem did.

('KWIC' OR 'KEYWORD') AND 'INDEXING'

'AUTOMATIC' AND ('THESAURUS' OR 'AUTHORITY LIST')

'RELEV. JUDGMENT' AND ('SUBJECTIVE' OR 'INTUITIVE')

WORK TABLE

	Doc. No.	Rel. No.	Opinion	Scientific	Info.	Literature
1.						
2.						
3.						
4.						
5.						
6.						
7.						
8.						
1.						
2.						
3.						
4.						
5.						
6.						
7.						
8.						

4.3 MARC File Search (CIMARON)

CIMARON is the name for a system of programs whose goal is to provide a facility within which the structure and behavior of large bibliographic files can be studied. The major impetus for designing and building such a system is described in detail in a previous report issued by the File Organization Project of the Institute of Library Research.* The primary motivation for CIMARON is that there is a critical need for a computer system with enough generality to provide the means to carry out studies of basic questions concerning the organization and access of large bibliographic files. Some of the basic questions, as studied by the File Organization Project, concern the nature of index keys and multi-level cross-reference techniques, the possibilities of compression encoding and decoding for disk storage, the major procedures for large scale data conversion to MARC II, and the relationship between the organization of file storage requests for retrieval.

From the point of view of library education, it was fortuate to have the availability of such a system, and the support for adapting CIMARON to the environment of the Information Processing Laboratory was readily provided. The result is that CIMARON is an interactive on-line MARC search system which has great value for library education as well as for library research. The capabilities of CIMARON are now such that they can support a broad spectrum of student and research activities, ranging from producing printed subject bibliographies for undergraduates to logging the searching protocol of users of the CIMARON system for the purpose of studying the search behavior of system users.

4.3.1 Program Description

The full description of the CIMARON program system is to be found in the File Organization Project reports cited earlier. A detailed and comprehensive report of both the commands and operating instructions of the CIMARON system programs can be found in The CIMARON System: Modular Programs for the Organization and Search of Large Files, a volume of this report. What is presented here is a summary of what the program is, what it can do, and what its current data resources are. This is done to provide a context for the subsequent discussion of the uses of CIMARON in various School of Librarianship curricula.

4.3.1.1 Data Base Content and Organization

Theoretically, any file of MARC II structure records can be used as a data base in the CIMARON system. There is no theoretical limit either to the complexity of the records or to the number of records in the file. Any format

*Cunningham, J.L., W.D. Schieber, and R.M. Shoffner, A Study of the Organization and Search of Bibliographic Holdings Records in On-Line Computer Systems; Phase I, Berkeley: Institute of Library Research, University of California, March 1969.

(not necessarily monographs) defined within the conventions of the MARC II structure can be processed by the CIMARON system, subject to practical limits of file size and disk storage of index files.

Currently, the CIMARON system has two files available for its users. The primary file is very large, consisting of 95,000 catalog entries, drawn from the MARC-form machine file data base of the U.C. Santa Cruz Library. This data base covers over 80% of the U.C. Santa Cruz campus Library system and is comparable to the resources of a large undergraduate library with some subject areas of special competence (e.g., Astronomy, due to the presence of the Lick Observatory in Santa Cruz). The second data base consists of 5,000 catalog records representing the holdings of the San Diego Medical Association. This is a specialized file dealing only with subject matter relevant to bio-medicine.

Theoretically, a data base in CIMARON can be accessed through any data element which is defined as a field within the MARC structure. A separate set of index generation and file linking programs within the CIMARON system do the work of extracting and organizing these linked index key files. Again as a practical matter, because of disk storage costs and limitation, the current CIMARON system supports the following major access keys to the two data bases mentioned:

Access	Santa Cruz	San Diego Med.
Author (Main & Secondary)	Yes	Yes
Title	No	Yes
Subject	Yes	Yes
Reduced Author	No	Yes

The "reduced author" access file refers to an algorithm (developed by Professor James Dolby) for reducing names to a vowel-less phonemic standard form in order to overcome orthographic ambiguity.

4.3.1.2 Major User Commands

The CIMARON system can be divided into two parts: file generation and organization, and on-line retrieval. While both parts have parameters and user controls, only the capabilities of the second section, on-line retrieval, will be described in this and the following sections.

Select Data Base. Any of the currently available CIMARON data bases may be identified and selected (e.g., Santa Cruz, San Diego).

Select Access File. Any of the currently defined index files may be selected (e.g., Author, Subject, Title). The index file selected is applicable only to the specific data base which has been chosen (e.g., Santa Cruz Subject, San Diego Title, etc.). Note that some data base/index file combinations are not legal (e.g., Santa Cruz Title).

Select Operating Mode. Two modes are possible: browsing and retrieval. The browsing mode enables the user to scan the CIMARON index files in order to develop and determine appropriate formulations for subsequent retrieval requests. The retrieval mode allows the user to enter formal retrieval requests which result in searches made against the data base and in the retrieval of appropriate records.

Enter Browsing Request. With this command the user specifies which alphabetic portion of the selected index file is to be displayed (e.g., Get 'Da Vinci', which might be a legitimate request for the Subject, Author, or Title index file). Only one index file at a time is scanned.

Enter Retrieval Request. This command is used to express search specifications to be carried out against the selected data base. Within this request three complexities are possible: first, the request may contain more than one term; second, search request terms may be connected by the Boolean operators AND, OR, NOT, AND NOT, OR NOT; third, more than one index file can be used. Example: (AUTHOR/'Da Vinci' OR Subject/'Da Vinci') AND Subject/'Anatomy'.

Process Browsing Result. When a portion of the selected index file has been displayed, the user may then advance the display any number of entries, save any term in the display, switch the display to the save area (i.e., list of saved terms), or print the display (or the save area) on the 360/40 line printer.

Process Retrieval Result. After the data base has been searched and some non-empty subset of records has been retrieved, the user may then display each retrieved record in the subset, print any record in the subset on the 360/40 line printer, switch the format of the record display from a user image to an unmodified MARC version of the record, or at any point, the user may terminate his examination of the subset.

Program Output. The direct result of the on-line interactive portion of the CIMARON system is an examination of either the index files (browsing mode) or a subset of data base records which correspond to a formal search request (retrieval mode). The results of browsing can be saved, printed, and used for further file searching exercises. Each index file entry contains a count of how many data base entries are related to that particular index file key. This is equivalent to giving the number of books indexed under a given subject, and can be useful in giving a sense of fruitful topics for search and retrieval.

The results of retrieval requests are full catalog entries with both descriptive and subject cataloging. If these results are printed, the retrieval request is printed also, thus permitting correlation and comparison of alternate search request formulations.

The CIMARON retrieval program operates in an on-line cycle that is part of the psychological "real-time" of the user. Thus data base searches, even under the worst conditions, (i.e., when all three terminals are in operation) take less than thirty seconds. The results are displayed on the Sanders CRT screens quickly enough to allow natural user responses and reactions. Thus

the system output is delivered with enough immediacy to stimulate further user requests in a cycle of request, examination, evaluation, and further searching.

4.3.2 On-Line Searching of Large Bibliographic Files

In a certain sense, CIMARON offers the simplest and most direct method of "exposing" librarians and library school students to computer processing as it may be applied to the most fundamental record in the library: the catalog card. For this rather large and important audience, CIMARON is an attractive introduction to computer systems because it performs natural processing functions (retrieval) against a familiar data base (catalog records). Finally, the on-line realtime operating cycle of CIMARON lends immediacy to the exposure, encourages interaction, and is free of the usual frustrations of batch-operated systems.

For example, consider the problem of errors, both of the clerical and syntactic type. All computer systems, from compilers on down, are notoriously unforgiving of even the mildest forms of error or ambiguity. This in itself is not a severe restriction either to teaching or to learning, except when it is coupled with a batch-operated system with three-to-four hour turnaround time. In that case the penalty for each set of errors, regardless of the magnitude or importance of the error, becomes an entire turnaround cycle, i.e., three or four hours. The burden on both students and teachers soon becomes intolerable, and the result is an overemphasis on the correctness of means and a corresponding feeling that the ends probably aren't worth it after all.

An on-line system in many ways justifies its increased costs by avoiding the situation described above. Of course user-generated errors still occur; however, the system's responses to such errors are immediate, and the terminal user (student or teacher) has the chance to consider what went wrong and to re-submit the transmission immediately without any significant delay. The terminals themselves facilitate such a re submission by offering a good variety of text editing controls such as insert and delete (see Chapter 3, section 3.6). Further, the programs attempt to check for improperly formatted or unrecognizable commands. Finally, even in the worst cases where the program or even the entire Terminal Monitor System may crash, reloading the system is only a matter of two or three minutes, and the Laboratory session can then resume.

Thus for the educator, CIMARON offers several strong advantages. It is relatively simple to describe the operating instructions of the retrieval and browsing programs. Since the goals of the system are familiar and obvious, one only need explain the means, not the ends. This is a simplifying assumption, and it makes it possible to present CIMARON to a class or group in a single lecture that can vary from fifteen to ninety minutes. The variation depends largely on the interest of the group in understanding the hardware configuration, the computer organization, linked file structures, MARC record structure, Boolean logic, and operating instructions.

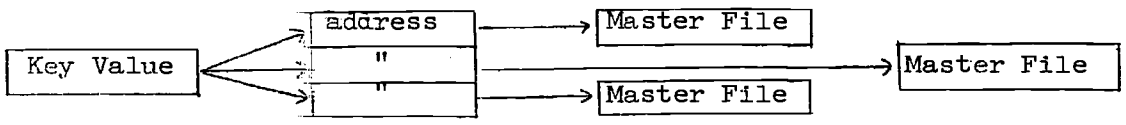
There are few enough modest size-cost on-line systems available. The Terminal Monitor System was built with about three man-years of effort, and runs on an expanded IBM 360/40 with one-half of a model 2314 disk as auxiliary storage. The cost of the three terminal hardware configuration is less than $40,000. From an economic perspective, the Information Processing Laboratory System is a reasonable prototype of what an on-line library system might be. Librarians and library students who use this Laboratory have the opportunity to experience and to evaluate what a realistic on-line computer system is. The system response time, reliability, operating principles, and the program protocols can all be experienced directly and exercised within an hour or two-hour session.

The educational opportunities here are especially rich. In a relatively short period of time, each terminal user can become a proficient participant in an on-line network. For most students and librarians this is a first direct experience with computing, and is likely to be their only chance to understand both first-hand and as individuals what a computer can and cannot do. In addition, the data bases that are made available by the CIMARON system are realistic for library operations. The files are bibliographic (catalog records, author names, subject headings), large (95,000 catalog entries), and represent the holdings of an existing academic library (U.C. Santa Cruz).

In fact real drawback of CIMARON is that it makes things look too simple. That is, the users in a brief session cannot understand the complexities of data conversion and index file control, and it is usually necessary to emphasize the prototypical aspects of the environment, the difficulties of an order-of-magnitude increase in file size, and terminal users which might be required for realistic library operations.

4.3.3 File Organization and Record Structure

From the point of view of advanced and sophisticated record and file structure techniques, CIMARON employs sophisticated techniques and formats, but these are not so esoteric that the general characteristics are unavailable to students with minimal technical background. For example, CIMARON uses a three-level addressing and file cross-reference scheme in which there is a master file, a key value file, and an address file. All three files are hierarchically linked in the following manner:



The key value file is interrogated in terms of the arguments of CIMARON search or browse requests. In its searching mode, CIMARON uses the address file to locate and fetch master file records which are stored in random locations of the 2314 disk.

For teaching purposes this can serve as an introduction to both the general notions of inverted file keys and the general techniques of file linking in computer systems. It also can be used to emphasize the crucial difference between manual and electronic systems, namely that the same master file record can be accessed easily by a number of different inverted topical keys (e.g., author, title, subject, series, publisher, date, etc.). In manual systems most frequently an extra copy of the master file record is needed for each additional access point which is created, which usually curbs the natural multiplication of access methods.

The topic of file linking and multiple access points also leads quite easily to the definition of the MARC record structure as the standard record structure which CIMARON accepts. File linking can be shown to be a phenomenon related to the leader-directory-data organization of the MARC record.* Further, the notion of multiple access composed of inverted file keys can be shown to be related directly to MARC monograph format definitions of fields and subfields. While CIMARON currently imitates the conventional methods of bibliographic access (author, title, subject), it is also possible to generate "non-conventional" index files such as series, publisher, subject subdivision, publication date, etc. In the near future we hope to have some of these access methods available for student use.

It turns out that CIMARON is not necessary for the presentation of the MARC II record format. Few librarians or library school students experience any serious difficulty in understanding the formal mechanism of the MARC format. The documentation (produced by the Library of Congress) is adequate and the contents of the format are familiar, even if the machine organization (leader, directory, variable fields, etc.) is not. However, it is still difficult to make the presentation of MARC as extensive as it might be. For example, it is sometimes a problem to demonstrate that a machine-form record such as a MARC form bibliographic record has a wider potential of applications than the manual form of the same record. There are as yet few major library projects which use the MARC monograph record to provide services which are strikingly different from the LC proof-slip service or the general card catalog. Even MARC based book-form catalogs usually parallel their card-from counterparts. Finally, and perhaps most seriously, it is difficult for students to dissociate the MARC monograph format from the more general definition of MARC as a standard record structure for all types of bibliographic records.

*See U.S. Library of Congress, Information Systems Office, MARC Manuals, 1969.

Thus if CIMARON is not needed to present the details of the MARC record, it is nevertheless a valuable tool for discussing concretely what MARC records are for, how they can be used in a context which is not like conventional card catalog searching, and especially how MARC can function as a generalized definition of a record structure within which it is possible to organize many different record formats (e.g., Serials, Maps, Phonodiscs, Abstracts, Personnel Files, etc.).

For this purpose CIMARON is considered as a general purpose bibliographic processing system. Both the on-line configuration and the cross-linked file structure are secondary topics, and are viewed as means to an end. The emphasis in presentation is on relating the field and subfield definitions of the MARC record structure to the index generation and record search aspects of CIMARON. The identification of subfields in a MARC format such as the monograph format is related directly to the problem of constructing inverted index keys based on that subfield. The function of bibliographic data element identification can thus be seen as a close correlate of the function of bibligoraphic record search and retrieval. This correlation provides insight into the needs and uses of data element identification at the field and subfield level. Without such correlation, the MARC monograph data element identification scheme will be perceived either as arbitrary or as a direct parallel of the Anglo-American cataloging rules. With the experience of CIMARON, it can be seen that format (as well as form) follows function.

4.3.4 Searching Protocols in Machine Bibliographic Systems

The students' first experience with CIMARON concentrates on deliberately paralleling familiar card-cataloging search procedures. Using the commonly found attributes of author, title, or subject as access points, students formulate requests to retrieve obvious subsets, such as: all works written by (or co-authored by) FREUD, SIGMUND; or all works about BACH, JOHANN SEBASTIAN. Through these simple exercises, students gain confidence in their own abilities to master the mechanics of the system's operation. They can also easily verify that CIMARON will produce predictable and accurate results, and so they build up some measure of confidence in the system. With such simple exercises students can also get a feeling of the system's speed and response time, and the ease of switching from one search request to another without having to traverse several yards of catalog card trays.

The next step is to proceed to non-conventional searching methods. A whole progression of these methods are available, and the following sections briefly describe the ones that we stressed in our teaching.

4.3.4.1 Partial Key Specification

The search logic of CIMARON performs a letter-by-letter match of the user's search request with the system index key files. Using the search request as the major control, whenever a complete match is found the search is considered to be successful, even if the index key is longer. Thus specifying "SMITH" as a search request will retrieve all the SMITH authors as well as SIMTHFIELD, SMITHSON, etc. This facility can be very useful when searching for a corporate author where the corporate subdivision is not known. The facility is less useful when the searcher wishes to focus on a narrowly defined search topic. With this slightly non-conventional search logic, students are presented with a novel capability which can not only be powerful but also precipitate an unwanted flood of results. CIMARON users are thus introduced early to the problem of the appropriateness of a searching tool relative to the nature of the searching operation.

4.3.4.2 Index File Browsing

Normally, it is impossible to get any accurate quantitative survey of the author or subject coverage in a collection. In CIMARON this facility is possible by means of "browsing" through the index files and requesting a display of the number of records indexed under a specific key (subject or author, for example). This count gives the number of titles in the collection written by a given author, etc. This quantitative data can be very useful in estimating the probability of successful retrieval operations, in avoiding zero-return and overflow searches, and in properly formulating search requests. Since these browser displays of index records do not provide the complete citation of works they represent, the searcher is led to consider the collection from a more formal and structural point of view than he might normally do. One limitation of this representation is the lack of an effective cross-referencing structure such as is provided by the L.C. Subject Heading categories of see also, see also from, see, and see from. It would be of genuine research interest if such cross-reference links were added to the file, plus a method for automatically chaining forward and backward through the associated terms.

In any event a quantitative survey of collection holdings has an important educational effect by providing numeric rather than textual cues for searching strategies. The browsing capability adds a new technique to bibliographic searching and request formulation, namely the interpretation of quantitative file data, and CIMARON offers the educator a way of teaching students how to make good use of such numerical data in organizing their search logic. With a new resource like this, the dependence on see also etc. references (which is crucial in manual systems) becomes far less urgent; therefore we have not made an issue of providing this capability in the files. Thus the student has the opportunity to learn a new approach to searching strategy which does not involve the use of a cross referencing structure.

4.3.4.3 Boolean Search Techniques

CIMARON allows three dimensions of search complexity. First, it allows search requests to include simultaneously many different search terms (up to sixteen are allowable). Second, it allows the user to combine different index attributes (such as author, title, subject, etc.) in a single request. Third, it allows the user to control the forms of search term and attribute combinations according to the rules of Boolean logic. This last feature allows the formulation of requests using the logical connectives AND, OR, NOT. Although Boolean logic is not an indispensible requirement of machine searches, the application of Boolean operators to search specifications for MARC files offers an ideal context for introducing librarianship students to the formal rules of manipulation for access to bibliographic data. From this, the students are able to develop skill in the fundamental professional practice of analyzing search specifications of any degree of complexity in terms of configurations of elementary logical relationships.

4.3.4.4 Non-Conventional Index

Currently, CIMARON supports access methods which closely parallel conventional card catalog searching methods. Both from the programming and from the instructional point of view, arriving at this level of stable system performance has consumed all the resources actually available in this phase of the Information Processing Laboratory project. However, the design of CIMARON extends into the next phase of the project, and it is this extension we wish to discuss now, even though these features are not implemented yet.

Specifically, we have in mind the generation of index files based on any field or subfield of any MARC structure record, including availability of non-monograph MARC-structure bibliographic data bases. Each (index) file represents some distinct attribute of the records in the CIMARON files. As the number of different attribute files increases, the meaningfulness of Boolean search combinations increases dramatically. It is trivial to request all books which are classified under the subject 'RUSSIAN LITERATURE'. However it is not trivial to request all books which have been published before 1920, have more than 100 pages, contain maps or illustrations, were published in England, and have RUSSIA or RUSSIAN appearing as a main heading or subdivision in their subject heading field. With the addition of more index files and more logical operators such as *greater than, before, after,* the possibilities of CIMARON searching become not only much more powerful than conventional card catalog search, but also more valuable as a teaching

A further point here is that as there are more possibilities for additional index files, the power to create these files also will pass the availability of non-monograph files in CIMARON increases. Users will have strong interests in the generation of index files as well as in their use in search requests. Educationally, this is important because it will lead the students even further into considering and controlling the basic foundation principles upon which CIMARON is based. It is also conceivable and hopeful that on-line construction of small scale index files can be made

available so that students may perform on-line experiments with the most critical aspects of record access and file generation.

4.3.5 Analysis of CIMARON Browse/Search Exercise

There are several courses in the School of Librarianship which make use of CIMARON. In Course 276, Survey of Library Automation, CIMARON Laboratory sessions occupy a crucial segment of the curriculum. Out of a thirty-lecture quarter, students are usually ready for CIMARON during the period of the 15th to 20th lecture after the midpoint of the course has been reached. The material covered prior to this point included: representation of information, encoding alphabets, computer architecture, identification of bibliographic data elements, structure of MARC II records, contents of MARC II records.

The primary role of CIMARON and the Information Processing Laboratory is to synthesize the diverse themes of the course: machines, data structures and bibliographic data elements. The synthesis is accomplished by encouraging students to interact with an operational machine system containing complex file and record structures with bibliographic content. By learning to manipulate MARC monograph records in CIMARON browse and search modes, a foundation also is given for the later segments of the course.

These later segments will include the task of defining the fields and subfields of a MARC format for some non-monograph data record. An important part of the sense of how to organize such a definition can be derived from the scanning and retrieval operations which are encountered during CIMARON exercises. The rules of CIMARON provide a framework against which to check the validity or utility of data element (field and subfield) definitions for non-conventional bibliographic records such as phonodiscs, dictionary entries, food recipes, printing collection references, theatre programs, animal tissue slide pathology collection, pet registries. These are examples of data topics selected by students in this course.

Following is the CIMARON exercise developed for Course 276. Its two parts, here represented as sections 4.3.5.1 and 4.3.5.2, include both browsing and searching operations. It is assumed that prior discussion based on the CIMARON Users' Manual, plus simple warm-up exercises, have established general competence for both terminal and program command operations. The first section of the exercise focuses on using BROWSER as a device for establishing valid terms to be used in later search requests. The exercise begins by asking a general question, "How do animals communicate," and requires the student to think of some possible search terms which can be used to retrieve a bibliography for this subject. The purpose is to underscore at the start the difficulties of using an unfamiliar indexing vocabulary, whether in a manual or an automated system. At this level, the simple "automation" of [illegible] intellectual procedures, such as selecting relevant [illegible], should be perceived by students as a myth. This kind of de-mystification is quite necessary, and frequently the limitations of machine systems will be emphasized in order to develop a balanced perspective in the relative roles of machine and librarian in automated bibliographic systems.

Another limitation of the CIMARON index files becomes apparent through this portion of the exercise, namely the lack of a cross reference or term association structure within the subject index file. During class discussion of the exercise, the concept of associative search (see Section 4.2) is briefly presented, and the notion of statistical co-occurrence is given as a possible aid for constructing and expanding search term lists. The feasibility and limitations of constructing term association files for MARC monograph records is explored in this discussion. (It also would be possible to reinforce this topic by constructing an exercise question which could be explored both in non-associated CIMARON files and also in a variety of association files available in LABSEARCH.)

The next two BROWSER questions attempt to bring students' attention to another unfamiliar limitation of machine searching: authority control. Even when the search term is already known (e.g., 'Libraries--U.S.'), the user must be aware of the many minor representation variations which can confuse the perception of literal-minded machine systems. Many authority control problems are deliberately left in the CIMARON index file in order to display graphically the results of differences in spelling, spacing, punctuation and the form of data elements (especially birth and death dates). This theme is tied back to earlier course presentations concerning the distinctions between the implicit recognition of data (the natural human habit) and the need for explicitness in machine representations of bibliographic records.

The last BROWSER question (tally 'University of California' corporate entries) restates this same problem and emphasizes the continuing need and the rationale for consistent standards of uniform bibliographic citation. After the completion of this portion of the exercise, the student should have a clear idea of:

- the representation of inverted key files, as utilized in CIMARON search operations
- the problems of translating search questions into index term vocabulary items
- bibliographic aspects of authority control
- the role of minor typographic variations in authority control problems
- how to characterize a collection in terms of quantitative index term counts

At this point the student is ready to proceed to the search and retrieval portion of the exercise.

The CIMARON exercise has two independent goals: first to use Boolean logic to compile a bibliography on 'Animal Communication', and second to use CIMARON as an analysis tool to investigate the multiple meanings of the subject heading 'Decision-Making'. Boolean logic is usually a new topic to library school students, and although it is presented and understood easily, there is a need for live experience to accustom students to the power and

convenience of using multiple search term requests and Boolean operators. Thus, students are requested initially to conduct searches one term at a time, following the conventional card catalog search procedures. Then the exercise allows students to reformulate the same request using a compound search expression with Boolean operators.

This part of the exercise demonstrates the following properties of Boolean requests:

- simultaneous use of more than one index term
- simultaneous use of more than one attribute file
- duplicate and redundant records are retrieved only once
- greater specificity is possible through the conjunction of search terms with *AND, OR, NOT* connectives.

As a final step, students are asked to construct a Venn diagram representing their search request.

The second question in this search portion of the CIMARON exercise emphasizes the concept of expanding search requests by using empirically associated index descriptors. Thus, if an author or a subject term (not specified in the search request) consistently appears in the retrieval subset, it is reasonable to assume both that some measure of associativity with the specified terms exists, and that the search can be expanded by means of this second order set of index descriptors. Students are thus led to consider retrieved records as sources for producing additional search requests. This notion also is carried out in the last part of the question which asks students to consider whether it would be useful to have additional access points (index files), for example, L.C. classification numbers, to include in search requests.

The final question in this set concentrates on the potential ambiguity or polysemy of index terms. A single term, 'DECISION-MAKING', is chosen, and an analysis is made of all the different contexts in which this term is used as a descriptor. Those contexts are: business, philosophy, jurisprudence, psychology, and international affairs. For each context, the student is asked to compile a list or cluster of terms which serve to separate the different usages of the term 'DECISION-MAKING'. Questions of overlap among the areas are asked, and the usefulness of Boolean search as a disambiguating device is presented. Finally, students are asked to consider the design for a procedure by which all such multiple-meaning subject terms could be identified. This entails the use of CIMARON searching as a system for organizing or examining all the entries in a collection by means of a common algorithm. At this point the on-line facilities of the laboratory operations can be used as an experimental basis for developing and verifying large-scale system design.

4.3.5.1 BROWSER Exercises

Below, in two parts, is the CIMARON BROWSE/SEARCH exercise.

1. (a) The general question is "How do animals communicate?" Think of some likely subject headings, and use BROWSER to browse around in the Santa Cruz collection. Write down likely subject headings you find, along with the number of titles for each heading in the Santa Cruz collection.

   (b) Are there any misspellings or typographic variations in the data you've looked at so far?

2. How many titles are there in Santa Cruz under the general rubric "Libraries—U.S." (ignore further levels of subdivision). List any spellings or typing variations you encounter.

3. How many titles in Santa Cruz have the "University of California" (any campus) as their main entry?

4.3.5.2 CIMARON Exercises

1. Select two or three of the most promising subject headings for the animal communication topic. First, enter some search requests using only one term at a time, then develop a search request using all of the search terms. Evaluate the usefulness of using multiple search terms and Boolean operators. Draw a Venn diagram of your request.

2. As you examine the set of retrieved records, use the H command to save and print records which seem promising and/or relevant. Are there any other access points (index files) you would like to have for this retrieval problem?

3. Examine the set of documents which has been classified under the heading 'DECISION-MAKING'.

   (a) Note the LC classifications and other subject headings which are assigned to these records. Make a list of the general subject areas which seem to be included under the term 'DECISION-MAKING', e.g., business, etc.

   (b) For each area in the above list, indicate the cluster of subject terms used to describe that area.

   (c) Is there any overlap among the clusters, i.e., subject terms which are associated with more than one area?

   (d) How can CIMARON be used to enable a user to specify the "kind" of 'DECISION-MAKING' he is interested in?

   (e) Describe a procedure for discovering other subject terms which may have multiple meanings.

4.4 Reference Search (REFSEARCH)

4.4.1 Methodological Innovation

In our view the Information Processing Laboratory can be used not only to study innovative methods of bibliographic access and organization, but also to make a contribution to many aspects of traditional librarianship. For example, Laboratory facilities can add the dimension of on-line response and realization to a field of traditional librarianship such as reference. This can be achieved by considering the Laboratory as an environment within which to consider reorganizing basic instructional procedures and finding new ways of thinking about a traditional subject matter. The availability of automated techniques offers encouragement to instructors who are seeking more powerful conceptualizations of subject matter for teaching.

To try out this approach in library education, we selected reference service as an area where it would be possible to experiment with methodological reorganization. Reference service has been a notable feature of public service in libraries for most of the century. The technique of reference service differs from more straightforward sorts of library practice in two particularly striking aspects:

a. The librarian makes no attempt to identify the contents of reference books in terms of conventional subject headings, except for one or two gross descriptors for the entire book. Instead, the contents are conceptualized in terms of categories of information, representing modes of discussion about general types of data.

b. In contrast to the usual library practice of providing general and public access to information, detailed classification of reference works is not customarily represented in any directly available device like a card catalog, but instead is "filed" in the memory of the reference librarian. The consequence is that effective access to the data depends upon a transaction sometimes referred to as the reference interview. There presently exists some healthy professional controversy about what ought to take place in this transaction, but it is generally agreed that this procedure lies at the heart of reference technique.

It follows then that training in traditional aspects of reference service must focus on helping students to master both reference work categorizations as well as the techniques of the reference interview. Thus in effect, students must be educated in the art of translating a patron's reference question into the information categories which characterize and organize the reference collection. The general procedure for this translation process is reasonably well understood, and teaching usually consists of lectures about groups of specific reference books, including the information categories to which they belong. This is followed by exercises consisting of reference questions, which focus directly on the books under consideration in the current instructional unit. This lecture-plus-exercises sequence is repeated until a sizeable corpus of reference books has been analyzed.

4.4.2 Reorganization of Reference Search Techniques

In the discussions with both faculty and student groups, three limitations of the above teaching procedures were printed out:

a. There is no standard way of eliciting or evaluating students' search strategies during exercises. The motivation is to answer a reference question accurately, rather than to formulate or evaluate a search strategy. There is no convenient way to record the progress of a reference search, so that both students and instructors may analyze the formal procedures which define the search strategy. This would, after all, constitute an explicit analysis of the problem solving protocols of the skilled reference librarian, and would be a valuable educational contribution to reference students.

b. The physical and visual shelf-arrangement of the reference collection tends to suggest limited categorizations of reference works. While physical arrangement usually provides a useful first-cut approximation to primary reference categories, it does not allow any possibility of faceted classification. There is a genuine need for a device that can be independent of static physical order, in order to show the rich, often hidden variety of sub-collections and services throughout the collection.

c. There is no means by which the student may search the entire collection quickly for likely sources of answers. There is no record of what a given reference work can provide except for syllabus sheets and Winchell. An occasional re-examination of the books themselves has its good points, but we would guess that it involves much floundering and many false starts. There is need for a device that retains data on the overall features and service of works and permits data interrogation at high speed. Having such a device, the student's main job would be to assess intelligently what to call for in the way of features and services, and to discriminate between works that the device said were alike, but that human eyes saw as quite different.

On the basis of discussion such as these, the staff of the Information Processing Laboratory developed an educational system called REFSEARCH to experiment with solutions to many of the problems expressed above. Specifically, we proposed a tentative methodology which would lead to better protocol analysis of the question-answering and search-organizing techniques which the experienced librarian brings to the reference interview. The core of the REFSEARCH system is therefore appropriately an indexing structure which attempts to provide students with an explicit framework of reference information categories. Within this index structure, there is developed a set of paradigms designed to explicate some of the formal categories relevant to answering questions.

The REFSEARCH system is thus a collection of procedures which attempts to model the reference situation in two directions. First, it reflects the ways in which a reference librarian may organize and categorize a specific reference collection. Second, it is a universe of discourse in which reference questions can be located and formulated. Thus REFSEARCH presents to students an instrumentality in terms of which they can understand the basic behavior of reference classification, the reference interview, question-negotiation, and searching strategy.

4.4.3 Major REFSEARCH Categories

The core of the REFSEARCH* system consists of three hierarchically organized categories used to analyze both reference works and reference questions:

a. Channel. There are 17 distinct categories (or nouns) which serve as main channels or entry points for describing major reference topics. Examples are: *Words, Natural Processes, Places, Products, Art Works, Persons, Corporate Bodies,* etc.

b. Modifiers. Subsumed under the basic channel or entry categories are some 81 adjective-type qualifiers, which serve to limit of and specify further the major category. Thus many of these 81 qualifiers are appropriate only for limited channel entry categories. For example, the qualifiers Real/Imaginary would apply to *Person, Places, Products, Corporate Bodies,* but not to *Words, Art Works,* etc.; similarly Slang and Abbreviations would apply to *Words* but not to *Persons,* etc. Two special sets of qualifiers allow the student to express specific historical periods of interest (e.g., 19th Century), or to express special typologies (termed subcollections, such as law, education, religion, etc.)

c. Service. The categories discussed in a. and b. above both relate to the content of the reference topic. This third REFSEARCH category expresses the kind of information which is being provided or requested. For example, reference questions about a Harvesting Combine can vary from

When was it invented?

How much does it cost?

Where is it made?

What does it look like?

*A full description for using the vocabulary and paradigms of the REFSEARCH system will be found in a separate volume of this report (Meredith, Joseph C., Reference Search System (REFSEARCH) Users' Manual). See also Meredith, J.C. "Machine Assisted Approach to General Reference Materials," Journal of American Society for Information Science, (XXII, May 1971).

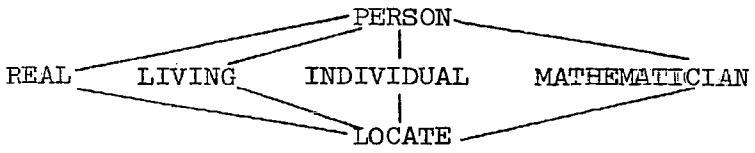
To accommodate this range of reference services (chronology, economics, geography, graphics), REFSEARCH provides nine categories of service: *Defines, Locates, Dates, Quantifies,* etc. Again, not all channel-service combinations are appropriate.

Thus the basic hierarchy is (1) channel (access point or handle) (2) modifier (qualifier) and (3) type of data (service). For example, characterize the reference question What is the mathematician Alfred Tarski's current address? The following 3-level organization could be used:

Level 1: (channel)

Level 2: (qualifier)

Level 3: (service)



The REFSEARCH language is used by students to analyze reference questions as general data structures. As an analysis of a reference-question, this is an idealization, but an important one. It is an accurate categorization of the reference question with a direct focus on the question rather than on the book in which the answer to the question can be found. This requires students to analyze the reference question more than to consider where the answer may be found.

The RESEARCH schema are thus a complement to other reference laboratory work, where the emphasis is on the reference collection in which the answers to various questions can be found. The REFSEARCH categories also are used in turn to analyze the general characteristics of reference works as these works are presented and discussed within the reference course syllabus. The students are able to ask both what kinds of services about what kinds of things are offered by a reference book, and in what ways a reference work can respond to a specific information requirement. Analyzing a reference work makes use of the REFSEARCH categories as a mirror image of reference question analysis. Instead of searching for the handle or channel required by a question, the student considers the channels, qualifications, and services provided by a reference work. Each reference work thus can be analyzed in terms of the general data structure categories it can provide as a resource for those questions which produce requirements for the same categories.

4.4.4 REFSEARCH Computer System

The specific task assigned to the computer and on-line terminals in Laboratory implementation of REFSEARCH is to bring together reference-question analysis and reference-work analysis as they are expressed in terms of REFSEARCH categories. This was done by means of the following procedure. First, the general reference collection of the School of Librarianship Reference Laboratory was analyzed in terms of the hierarchical REFSEARCH access, qualifier, and service categories. This collection consists of 160 titles (790 volumes). A few general purpose encyclopedias were excluded from the analysis.

The results of this analysis were then encoded as a machine-readable file in which each reference work was an independent record. The data elements of these records reflected the presence or absence of the REFSEARCH access points, qualifiers, and services provided by each work to a potential searcher. This preliminary analysis, was carried out by Laboratory staff in consultation with faculty members of the School of Librarianship. The machine-file version of the analysis was keyed and loaded and made available to a searching program which operated in an on-line mode under the Laboratory Terminal Monitor System.

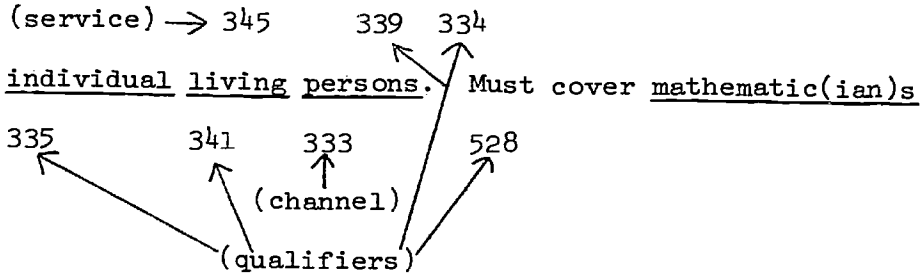
The REFSEARCH computer program component thus operates interactively using the Sanders video terminals, and has as its basic task the matching up of search requests with those reference works which can provide the channels, qualifiers, and services which are required. The search program requires no hierarchy or sequence of REFSEARCH vocabulary; REFSEARCH terms can be entered in the order most natural to the reference question. No Boolean connectives are used between search terms.

A person wishing to use the system enters code numbers corresponding with the terms listed in the classified index of channels/qualifiers/services furnished in the REFSEARCH Users' Manual. Queries are submitted as strings of code numbers with comma separators. Invalid codes are detected and displayed for editing. Retrieval on an acceptable request takes 2-4 seconds, and output is in the form of a display indicating the total number of works retrieved, listing their titles in accession number order. The display also repeats, in columnar form, the components of the student's original encoded request.

To illustrate the foregoing, let us use the example of finding a likely source of Alfred Tarski's address.

The specification is annotated with the numerical code corresponding with each element:

Required: a work which locates real proper-named



The student enters the codes at his terminal in the form:

345,339,334,335,341,333,528

and sends it to the computer. Within four seconds the following display is given:

```
LAB REFERENCE WORKS SEARCH PROGRAM

6 DOCUMENTS SATISFY THIS REQUEST
345     060   LANGER. ENCYCLOPEDIA OF WORLD HISTORY
339     069   AMERICAN MEN OF SCIENCE
334     109   INTERNATIONAL WHO'S WHO
335     110   CURRENT BIOGRAPHY
341     114   WHO'S WHO IN AMERICA
528     116   WHO'S WHO IN THE WEST
```

4.4.5 REFSEARCH Exercises

We retained the practice device of assigning student exercises in the form of typical reference questions. But now, instead of placing the stress on "finding the right answer," we shifted the emphasis to the problem of analyzing the questions in terms of a set of formally defined properties which are available in the REFSEARCH hierarchical categories of channel, qualifier, and service.

Students were asked to make explicit analyses of reference questions by identifying the individual elements of the questions and assigning those terms in the REFSEARCH language. Students thus, in effect, were asked to

transform reference questions into a configuration of formal search requirements, which were then input to the REFSEARCH program via the Laboratory's on-line terminals. The REFSEARCH terminal system analyzed the search specifications and provided an immediate listing of the books which would satisfy the request as formulated. Students then evaluated the resulting list and reformulated their searching strategy, if necessary. If the list was too long, then more specificity was added to the request. If the list was too sparse, then some of the search constraints were relaxed. Students recorded these changes of strategy, and the strategies were later reviewed by laboratory staff and faculty members.

4.4.6 Sample REFSEARCH Exercises

Four examples of this process of developing a strategy are given in this section. The examples are taken from a student's exercise, produced during the 1970 Winter Quarter. The term "unit" or "group" which appears in the student's discussion of search strategy refers to a syllabus unit used in the Reference course. The comments are the student's own evaluation of his analysis protocol.

Each of the exercises illustrates some interesting aspect of student usage of REFSEARCH. The first exercise, *Who is the President of the Senate of the New Mexico Legislature,* shows a shift in the channel access method chosen, from personal name to corporate body. This can be seen as an altering of perspective on the question from *President* (sentence main noun) to *of the Senate* (modifying clause). In the second exercise, *What does "privattryk" mean in the imprint of a Norwegian book,* three different access channels are used: products, words, and press. The last is most useful, but produces too large a list. With the addition of the qualifier, foreign terms, the result is a satisfying single appropriate title: Language of the Foreign Book Trade.

Exercise number 3, *What salaries do members of the U.S. Commission of Fine Arts receive,* is somewhat less satisfying. After three tries, no reasonable retrieval list emerges. As a result of the regular reference curriculum, the student knows the answer beforehand and uses the exercise to evaluate his formulation of a strategy. A similar experience occurs with exercise number 4, *Who is in charge of public relations in Niagara Falls, New York.* The notable occurrence here is that an unexpected candidate is retrieved, but passed over in favor of a known result.

Several features of this process are noteworthy. First of all, just as in an operating reference situation, there are a variety of reasonable analyses of a reference question, each of which may lead to a somewhat different set of answer-providing sources. The interactive capability of the Information Processing Laboratory provides the student with a convenient and rapid method for successively submitting multiple interpretations of the same question and comparing the results. Aside from the obvious benefits of helping inexperienced students to develop an understanding of how different interpretations may affect the choice and quantity of answer-providing sources to be selected, this also helps the student to regard the analysis of a reference question as the specification of prospective search criteria.

In this connection it is of interest that no attempt was made to urge students to memorize the set of index characteristics that had been assigned to the individual reference sources in the data base. Although significant properties of these titles were discussed in lectures, and students were free to consult their class notes in working out their question analyses, the primary emphasis in the exercises was on the details of the analytical process rather than the descriptive features of the books.

EXERCISE NO. 1: What is the name of the president of the Senate in the New Mexico state legislature?

	1st Request	2nd Request
Student Response:	334 proper name	433 corporate body
	335 individual	436 real
	336 capitalized group	434 proper name
	339 real	444 key person
	431 living	518 government
	343 identifies	558 USA
	518 government	644 1968
	558 USA	
	642 1970	

Student Comment:

First attempt retrieved no documents. Evidently it was either too restrictive or there was an element of conflict involved. The second attempt produced a list of thirteen documents of which six were in booklist group 11. They were: Statesman's Year Book, Europa, Year Book of International Organizations, U.S. Government Organization Manual, Official Congressional Directory, and Book of the States. The last name was chosen as most appropriately offering the answer. The name E. Lee Francis was found in the 1968-1969 edition on page 553.

EXERCISE NO. 2: A book published in Oslo has "privattryk" in its imprint. What does this mean?

1st Request: 359, products, common name; 360, real; 363, individual; 365, identifies; 521, press; 561, multi-national.

First request retrieved only the World Almanac.

2nd Request: 210 words, foreign terms; 212, nick/real name; 202, defines; 521, press; 561, multi-national.

Second request retrieved no documents.

3rd Request: 521, press; 561, multi-national.

Student Comment:

Third request produced 15 documents, among them Language of Foreign Book Trade from group 4. A fourth attempt was made by adding 210, foreign

terms, to the formulation of the third attempt. This retrieved one document, the Language of the Foreign Book Trade. The listings are alphabetical by language, and the answer is found on page 33.

EXERCISE NO. 3: What salaries do members of the U.S. Commission of Fine Arts receive for their services?

1st Request:	434 corporate proper name	518 government
	436 real	540 plastic arts
	447 non-profit	558 USA
	445 $	

Student Comment:

This listed ten documents, with only Year Book of International Organizations from group 11

2nd Request:	434 corporate proper name	445 $
	436 real	518 government
	438 identified	558 USA

Student Comment:

This gave one less document, but three from group 11, Statement's, Worldmark, and the above Year Book of International Organizations. A third attempt was made purposely less restrictive in hopes of including U.S. Government Organization Manual in the list.

3rd Request:	433 corporate bodies	445 $
	436 real	558 USA
	434 proper name	

Student Comment:

This expanded the list all right to 21 documents, but produced only the same three from group 11. Adding 518 (government) to the input of the third try cut the list to 14 documents, but still included the same three only from group 11. One would be willing to bet that the combination of:

434 corporate proper name	518 government
436 real	558 USA
447 non-profit	
439 discusses	

would include the desired document. At any rate the answer is found in the U.S. Government Organization Manual on page 430 that members serve without pay.

EXERCISE NO. 4: Who is in charge of public relations in Niagara Falls, N.Y.? Give both his title and his name.

1st Request:	302 place proper name	545 tourism
	304 real	558 USA
	518 government	334 person proper name
		337 capitalized role
2nd Request:	434 corporate proper name	444 key person
	436 real	518 government
	438 identifies	558 USA
		667 1969-date
3rd Request:	302 place proper name	518 government
	304 real	558 USA
	444 key person	667 1969-date

Student Comment:

> The first attempt produced first one document, World Almanac, which seemed certain not to have the answer. The second attempt produced no documents at all. The third attempt produced two documents not in group 11, Keesing's Contemporary Archives and Congressional Quarterly. Without study it was not known whether Kessing's might have the answer. In anticipating the answer to be in Municipal Year Book, the three requests were redone dropping, one at a time, codings which might possibly be poor judgment in regard to the question. After some nine or ten attempts which would not produce the document desired, the effort was abandoned. The book, as anticipated, does produce the answer by use of index and several separate elements of information. The 1967 edition is undoubtedly out of date, however. Page 259 and page 261 indicate that the mayor fulfills this function. And page 605 names the mayor, E. Dent Lackey, whose term expired in December, 1967.

4.4.7 REFSEARCH Evaluation

The REFSEARCH Laboratory system was described in the following way by a member of the School of Librarianship reference teaching faculty:

> The students have specific questions for which an advanced strategy has been worked out and tested. They are instructed to devise strategies designed to elicit titles of specific works from the computer system, or strategies that result in the titles being included in whatever is returned. They are to report the strategies used and the titles returned in each instance. Their experience then can be compared with the worked-out strategy. I think this controlled use of the program will prove valuable for discovering how students are thinking about how to answer a specific question.

The fact that the system produces a list of a set of titles in response to a specific strategy marks a significant departure from more traditional

methods: when a manual search of a set of reference books is performed, the answer to the question may be found in the first source selected. The REFSEARCH system responds in terms of document sets, and the result is wider student exposure to a number of reference books, as each individual source turns up in a different context. Correspondingly, a work which overlaps another in many ways, and seems to turn up as its companion in most retrievals, becomes conspicuous when it fails to do so, thus drawing the student's attention to important differences between the two works.

A serious question in the REFSEARCH system concerns indexing the collection; that is, encoding into REFSEARCH terms the features of the reference books chosen for our data base. How does a person index likely or "surprise" features of a reference tool? If the reference works in the collection are coded only according to their broadest and most obvious characteristics, the system consistently will retrieve plausible choices of answer-providing sources in response to accurate analyses of input queries. However, this reliability is achieved at a considerable loss of intellectual and bibliographic nuance. The retrievals are too predictable and fail to reflect the important professional technique of making strategic use of reference tools for purposes other than those for which the tools are primarily designed. For example, the Statistical Abstract of the United States and the Worldmark Encyclopedia contain many useful literature citations, but they are not primarily bibliographic tools. It would be desirable to take account of their potential utility as bibliographic sources in the indexing, but coding them as bibliographic works leads to the unfortunate results of producing too large and indiscriminate an output in response to queries for which sources chiefly concerned with bibliographical data are desired.

In our original indexing of the data base, we were generous in our coding of these surprise features, and failed to anticipate the number of noisy retrievals that this approach would entail. This sometimes produced outputs which students recognized immediately as being contrary to common sense, and they unfortunately tended to regard this as a failing of "the computer" rather than attributing it to the dilemmas of the human beings who had made the indexing decisions.

4.5 Machine Tutorial Mode (DISCUS)

4.5.1 Introduction and Summary

In the Information Processing Laboratory project, the term Machine Tutorial Mode (MTM) is used to define a special form of Computer-Assisted Instruction (CAI). As we stated in our Phase I Laboratory report, MTM connotes free and flexible dialogue between a teacher and an adult student, especially tailored to the needs of graduate instruction in library science. Thus it does not fall into the same category as other forms of CAI.

The most salient characteristic of MTM is the ability to engage the student in an active conversational interchange. Exposition is accompanied by frequent questions and other opportunities for the student to express himself, and these responses will, in most cases, directly affect the unfolding of the presentation. This dialogue form of presentation requires a computer system with a sophisticated branching ability, so that the presentation of material can be varied according to student responses. Program reactions need to be varied to suit categories of responses - commending those that contain desired elements and supplying corrective instruction in response to those which are in error. In addition, an MTM system must be able to record student performance in detail, for both individuals and groups. This ability permits not only the evaluation of student performance but also refinement of the course itself.

In order to develop a CAI system embodying these characteristics, however, one must be prepared to invest heavily in terms of overall system development, as well as in the writing of specific courses. Thus during this phase of the Laboratory project the most important single technical development relating to work in MTM was the design and implementation of the DISCUS System. DISCUS is designed to support programming and course writing activities specifically tailored to the MTM requirements outlined above. DISCUS is an interpretive system designed to operate with the on-line consoles available in the Information Processing Laboratory in Berkeley as well as with the terminals available in the URSA time-sharing system at UCLA (both systems use IBM S/360). The DISCUS system is equipped with functions which carry it well beyond the range of problems normally associated with CAI, and was jointly sponsored and supported by an additional ILR Project.*

4.5.2 DISCUS and PILOT

Two factors led to the decision to design and develop DISCUS: (1) the core requirements of the PILOT CAI language** prevented its being implemented for the IBM 360/40 system on the Berkeley campus, and (2) PILOT interfaced only with teletype and typewriter terminals, while the Information Processing Laboratory had cathode-ray-tube terminals available for CAI work. The textual presentation crucial to MTM work in fact depended heavily upon CRT speeds, i.e., instead of being limited to one or two sentences, displays often ran on to two or three paragraphs - far too

*OEG-1-7-071083-5068, File Organization Project.
**Karpinski, R., et al., PILOT (Programmed Inquiry, Learning or Teaching) User Guide 12-1-68: a Conversational Computer Language, San Francisco: Office of Information Systems, University of California Medical Center, 1968.

much for the speed of a standard mechanical terminal. Our early work in PILOT was well worth while, as it permitted us to specify precisely the characteristics required in a CAI language tailored to the Information Processing Laboratory and to the kind of CAI programming to be accomplished in that environment. PILOT is a versatile, "powerful," easily-learned language, and nothing in this report should be construed as deprecating it as a CAI medium. The capabilities of PILOT were in fact basic to the specifications which we laid down for DISCUS, but with the added imperative that DISCUS operate within a small partition of core memeory and that it interface with CRT terminals.

DISCUS is distinguished from PILOT and, as far as is known to us, from most other CAI languages by its use of a block structure which stratifies processing at various levels within a given frame. In other words, instead of a program reaction to a given input being dictated by a single condition-code setting, it can be shaped by a pattern previous success or failure conditions. This kind of CAI programming, which is essentially two-dimensional rather than linear, can become quite complicated in meeting a complicated objective, or can proceed very simply and directly. The design reflects a policy quite different from that which underlies CAI language limited to a half-dozen or so simple commands that in reality represent fixed subroutines.

DISCUS was coded entirely in assembly language, and is very fast, even though it is essentially an interpretive system. This is true for both the compile and execute times. Core requirements for both compilation and execution are small enough to make implementation in a small computer eminently feasible, and for this reason we feel that the system will be well suited for use in schools having limited computer resources. A detailed description of the DISCUS language and system is contained in an independent volume of this report, DISCUS Interactive System Users' Manual by Steven S. Silver and Joseph C. Meredith (Berkeley, 1971).

DISCUS has been available for use in the Information Processing Laboratory for about a year, and in that time has proved itself reliable in the execution of the one operational CAI program which has used it, namely the Subject Cataloging Course. A program which automatically translated the PILOT version of the course to DISCUS was produced and this facilitated the shift from PILOT to DISCUS for both the above course and its supplement, although the latter has not yet been operationally tested.

New coding in DISCUS has consisted of experimental routines designed to test various capabilities, and ongoing work in connection with the Subject Cataloging Supplement course. This last effort has been carried forward under the guidance of School of Librarianship faculty.

4.5.3 Potential Applications

While our earlier commitment for MTM was in the area of Subject Cataloging, the expanded capabilities of DISCUS invited consideration of other subjects which might lend themselves, in part at least, to computer assisted instruction in a machine tutorial mode. One approach is to work in terms of short self-contained subject areas, each representing a fairly

discrete aspect of material that would be suitable for MTM techniques. An example of such an area might be catalog card formatting, which is usually taught in conjunction with Descriptive Cataloging. This would be a simple and effective MTM sequence, involving generous student participation. The MTM sequence could be used independently by students, or in conjunction with a formal cataloging course or laboratory.

The selection and advocacy of new applications of MTM techniques needs to be carried forward in an atmosphere of mutual understanding and enthusiasm between the MTM programmer and the faculty member concerned. The former needs to be fully informed about course objectives and curriculum policy. The faculty member should be able to assist in exploiting the computer's potential for educational support in library science.

While none of the applications studies noted in this section have advanced beyond the preliminary stage, they are nonetheless worth reviewing briefly as potential contributions of the Information Processing Laboratory to innovative educational techniques.

4.5.3.1 Potential MTM Applications to Library Administration

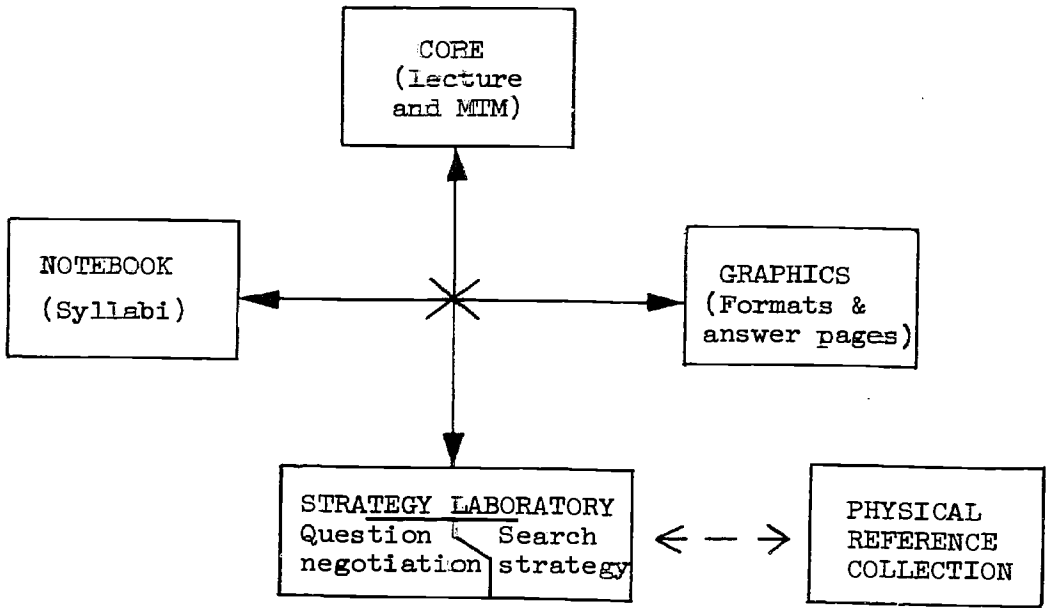
Course work in this subject involves detailed consideration of administrative structures and conscious attention to principles so well ordered that they lend themselves to a kind of check-list treatment. It also includes engineering and architecture where these bear upon problems of traffic flow, storage, and other aspects of library design. Some of this material could be presented for demonstration, analysis, or review very conveniently through one or more MTM courses integrated with lecture material using well designed accompanying graphic aids.

Textual material could cover some of the braod outlines of the subject, most of which is quite straightforward. With this in mind, we asked the senior faculty member who teaches this course to recommend material from the textbook currently in use. A brief resume was coded, and the result was put aside as a possible starting point for future development.

4.5.3.2 Applications to Materials in General Reference

The teaching of general reference skills was proposed early in the project as the next most likely application of MTM after Subject Cataloging. MTM was seen first as a drill-and-practice medium which could streamline the absorption of the bibliographic and functional details of standard reference works which students had to memorize in order to "perform" well in the Reference Laboratory.

The faculty member chiefly responsible for the teaching of the General Reference course in the School of Librarianship enthusiastically supported our interest in developing some kind of MTM tool for enhancing instruction in this vital segment of the curriculum. The following structural concept was proposed:



There seemed to be an implicit need in the proposed design for the development of a new typology for organizing reference works, both from the point of view of teaching and practice. A different approach to the teaching of reference librarianship would hopefully give students a better insight into the nature of general reference data while making fewer demands on pure memory. In order to de-emphsize the study of individual works as separate, monumental entities to be "mastered," we attempted to identify principles which would describe the collection as a whole, both in terms of data embedded in the collection, and in terms of the network of paths to the data.

This effort became the "REFSEARCH System" which is discussed at length elsewhere in this report. It is interesting to note that in this instance the Information Processing Laboratory played a major role simply by being there, i.e., by placing certain equipment and software (including MTM) at the disposal of faculty and technical specialists.

4.5.4 MTM Course in Subject Cataloging

The MTM Course in Subject Cataloging is an introductory course of instruction in the principles of alphabetical subject cataloging as represented by Library of Congress practice. The course is intended to give a beginning student of librarianship an appreciation of some of the factors which influence the cataloger in his choice of subject headings for various works, and an understanding of the network of references inserted in the catalog for leading the patron from terms not used to the terms which are. The discussion prepares the student for practical exercises in the cataloging of real books, according to LC practice and (to a lesser degree) according to the abridged practice advocated by Sears for small libraries. Extended discussion of difficult choices is avoided, the course being designed primarily as a review and reinforcement of lecture presentations. However, it is not keyed to any particular class schedule, and could thus

be used independent of formal curriculum. The course does not include any extensive material on classification, on classified subject cataloging, or on descriptive cataloging.

The study of subject cataloging troubles many students because cataloging practice in many respects seems to be inconsistent where consistency is most needed, and seems to be illogical where logic would seem to point to single, simple solutions. The teacher of subject cataloging needs to advance various precepts on the one hand while explaining why they don't always work on the other. Thus a considerable amount of interpretation is necessary in order for the student to deal realistically with the type of cataloging problems which he must expect to encounter professionally.

The MTM Subject Cataloging course is organized around two related concepts: the way people think of things and ideas, and the way things and ideas are embodied in books. The content of a document is the sum of the ideas or data which it contains. The problems of identifying a document and of describing its content in such a way that a person other than a librarian or an author may find and use it are the concerns of the bibliographer and of the librarian. Identifying the contents of the document denotes a unique position in a specific bibliographic universe. Thus the assignment of an elementary "subject term" combines references to a great many documents under a single description, and also refers to a specific grouping or bibliographic frame of reference.

A compendium of these groupings, combined to form a subject catalog, provides intellectual access to documents for users. It attempts to provide a kind of retrieval service through describing with the utmost brevity the content of documents. Fortunately, the subject heading structure is shored up by the bibliographic, or so-called "descriptive" entries attached to it (which are, of course, identifactory as well), by the author/title catalog (with its parallel descriptive services), and by the classification-number/shelf-number apparatus (which has some descriptive qualities too). The catalog as a whole, then, provides the user with a range of access and descriptive services, clustered in two different ways: explicitly according to content (in the subject catalog) and explicitly according to origin (in the author/title catalog).

The content of "Course 210X" (Subject Cataloging) touches upon all these matters - obliquely, to be sure, but in ways which we have tried to make consistent with the underlying philosophy. First, the student is confronted with the problem of representing a collection of documents in such a way that a literate person may approach a subject through a variety of terms and be led to clusters of related material on that subject. The first third of the course is occupied with this objective and the modes of furthering it.

This leads to consideration of the idea of subdivision as a control device. The four types of subdivision are discussed, and the student is cautioned against their use in a proto-classificatory way inconsistent with the idea of dictionary arrangement.

Then the student is introduced to the "Sears List," less for its own sake than as a means of reinforcing his perception of LC policy, by directly comparing it with Sears' more coarse-grained treatment. Finally, special

situations are touched upon, such as the effect of ethnic factors in the formulation of "literary headings." This concludes the course.

The material is presented in 253 frames of varying length and intricacy, joined by logical connectives (cause/effect, elaboration, example, contract, exceptions, ramifications) designed to lead the student's attention continually onward. Each frame begins with a statement, followed by a question. If the student answers correctly, execution passes to the next frame; if incorrectly, the program supplies a cue and repeats the text and question.

In order to accommodate student responses which the MTM programmer failed to anticipate, a "catch-all" response is provided at the end of each frame. It is worded in ways calculated to sound fairly responsive, but not deceptively so (e.g., "I don't understand what you mean by (the student's input) in this context. Please re-word your reply.").

Although the course is not outwardly segmented, the presentation does move successively through certain broad areas of interest indicated in the preceding section. The ordering of individual topics in these areas is described in the Subject Cataloging Course Outline, given in Section 4.5.6.

The style of presentation is varied, but tries to adhere to a strong narrative line which will capture and retain the attention of a highly diverse set of students. Didactic statements are interspersed with some of a more speculative nature. We believe that textual statements should be models of clarity and precision, and should err, if at all, on the side of saying too little rather than too much. Then, if the student fails to grasp the idea of the statement, that fact will be apparent in the ensuing interchange, and the program will supply correction and clarification accordingly.

The student is aware from the start that the computer cannot converse with him in human fashion. We believe that he should be brought to realize, however, that a human teacher is trying to communicate with him through the computer medium, and that the progress of the interchange can be dynamic and exceedingly variable. The conversational tone is heightened by using the first person singular instead of plural; the student should never be given the feeling that he is dealing with a committee. Calling the student by name, or referring to something he said previously and relating it to the corrective or reinforcing matter to be conveyed are both good devices. Flat rejection of a student's response is avoided; instead, he should, if possible, be told wherein he has erred. The responses designed to handle unanticipated answers need to be especially tactful, for all too often some of these answers are quite reasonable.

4.5.5 MTM Subject Cataloging Course Supplement

The MTM Subject Cataloging Supplement (201XL) is a program that provides computer guidance in the assignment of subject headings to real books located in the Information Processing Laboratory. Interactive conversation is limited mainly to discussion of the book in hand, which may be any one of the thirty-five works that comprise the "collection." No single student would ever be required to catalog all 35 books in the program, but we estimate that if this were attempted by a reasonably skillful person it

would take about nine hours. Actually running time varies radically according to levels of student participation and skill. The speed differential between CRT and typewriter (or teletype) terminals is less striking than in the case of the basic course, because displays are generally brief, and a lower percentage of terminal time is spent in actually typing and receiving messages.

The objective of the Subject Cataloging Laboratory Supplement, as with conventional laboratory arrangements for students of subject cataloging, is to afford students an opportunity to put into practice with real books the rules and principles they have learned through lecture, reading, and the Basic Course described in the preceding section. One of the necessary conditions for proper cataloging of a book is to go beyond the title page and title page verso to try to make an independent estimation of what the book is about. For exercise purposes there is no substitute for the book itself. Dealing with real books also brings a sense of immediacy to the subject of cataloging that can be obtained in no other way. It reinforces that which the student has learned, while at the same time drawing his attention to elements requiring further study.

In conventional cataloging laboratories the student's work is usually structured around a subset of works representing a certain problem. Students are required to choose ten or so of these and devise subject headings for them. (This exercise may be combined with the assignment of classification numbers, formulation of descriptive materials, etc.) One drawback in this arrangement is that the laboratory assistant's attention must be divided between twenty to forty individuals, and the student seldom receives an evaluation of this work until it has been handed in, checked, annotated, graded, and returned. The procedure can result in the student's not knowing for several days whether his choice of subject headings was suitable, during which time his reasons for having made these choices may escape him, and thus - in an educational sense - any remedial commentary loses force.

The principal advantage to be gained from computer-administered laboratory exercise is that (as in ordinary MTM) the computer responds immediately to input from a student terminal. When students are able to receive immediate feedback on their choices, they can quickly revise and resubmit them until they achieve the desired result. Also, they can carry on a discussion of various choices without the feeling of finality which builds up around a written laboratory assignment. The actual books used by the Supplement were drawn from the Library School's regular cataloging laboratory, whose staff was helpful in furnishing preferred headings and permitting us to review student returns in order that we might establish the range of fallacious choices that the machine program should be prepared to deal with.

In the main, the program encourages a simple, logical approach, in order to counteract a frequent tendency on the part of students in a laboratory situation to mistrust obvious solutions. Since it would be undesirable for a student to tie up a terminal while not actively engaged in an interchange, we propose rotating one terminal's use among three students.

A student should spend at least twice as much time in assessing the nature of a work, consulting the LC List, and writing down suitable heading(s),

as he spends in discussing his choice(s) with the machine program. Accordingly, we structured the operation in terms of three students per terminal, working in whatever rotational sequence would be most comfortable to them as a team. There is much to be gained through this type of arrangement, as we have since found out in connection with work in other contexts.

Within the above framework, student computer interaction proceeds as follows:

a. The student is asked to identify the book in hand. (Some misspellings are accepted here.) The program then displays full title and author, which the student confirms if, in fact, it is the one he wants to talk about.

b. The student is then asked to submit his choice(s) of heading, one at a time. If recognized, they are either accepted or declined as unsuitable and in the latter case the reason for their not being considered appropriate is given. If they are not recognized, the student is asked to formulate them in some other way, check his spelling and punctuation, etc. (Some, but not all, of such errors can be anticipated and provided for.)

   It is fairly difficult to detect "noise words" in input to a CAI system as open as DISCUS, but if the student conscientiously adheres to the headings given in the LC List, such extraneous material will not affect the acceptance or non-acceptance of a particular subject heading term.

c. The program encourages a student to stay with his problem until it is solved. If he says something like "I give up" rather too early in the game, he is told to keep trying. If he asks for help, he gets some appropriate suggestions. If he continues to flounder, he is given the correct answer(s).

d. The counters which govern the above features are fairly well protected from redundant input, which could otherwise deceive the program into crediting the student with more correct answers or completed books than is actually the case.

KEY TO TYPES OF QUESTIONS	NUMBER OF TIMES OCCURRING
Q1 True, false	5
Q2 Yes, no	24
Q3 Fill in word(s)	37
Q4 Multiple choice	59
Q5 Match list items	5
Q6 Fill in phrase or statement	12
Q7 Formulate heading(s)	23
Q8 Give examples	1
Q9 Give an opinion	1
Q10 Give unconstructed answer	14

4.5.6 Outline of MTM Subject Cataloging Course

201X: INDEX

SECTION ONE

Label and Q-type	Frame Topic Summary
	A. THE FUNCTION OF A CATALOG
Alice (No question)	Variety of readers, terms for topics and the variety of document language make catalog construction complicated.
Abarca (Q4)	The primary role of a catalog is reader-access to books, in contrast to: (1) reference function - mainly for staff use, (2) inventory function - the shelf list.
	B. CATALOG AND SHELF ARRANGEMENT
Priscilla (Q3)	Catalog and shelf arrangement contrasted as retrieval devices. Limitation of shelf arrangement (1) several topics but only one physical location possible.
Abauzit (Q4)	(2) A book not on the shelf may be either circulating or not in the collection - no way to tell for sure without a collection record, such as the catalog.
Absalon (Q3)	Terms distinguished: book and document
Abbatucci (Q3)	The card catalog as a retrieval device.
	C. DICTIONARY AND DIVIDED CATALOG
Abbe (Q3)	Dictionary cataloged characterized. Interfiling of entries and alphabetic arrangement.
Frumpy (Q4)	Dictionary catalog may be consulted without an index. User-search success a function of cataloger's skill in selecting and formulating headings.
Abdalah (Q4)	Formulating multiple headings for books often with several topics results in a complex subject heading network.
Marian, Abdelaziz, Abdul (Q2)	Dictionary catalog presents a size problem to which the divided catalog is a solution.

Label and Q-type	Frame Topic Summary
	D. MULTIPLE AND DIRECT ACCESS
Abenezra (Q3)	Multiple headings for a book provide multiple access.
Screech	Multiple access combines with direct access provided by selecting the most widely used term for the main heading.
Abercrombie (Q6)	Selection of main heading term. The one giving maximum direct access from among several candidate terms.
Abernathy (Q6) Thorpe	Terms not selected as main headings are used with se references to the main heading.
Abert (Q6)	Example: Use of "great silver bird" on an Indian reservation. Haykin-"the reader is the focus. . ."
Abich (Q2)	The main heading as a "majority" term often becomes standard for most libraries. Alternate terms often reflect the interests of the local user population.
Aboo (Q2)	Another approach of listing books under all terms used to describe a topic might result in a chaotically huge catalog.
	E. MULTIPLE TOPICS OF COMPOUND SUBJECTS
Accolti (Q4)	A book with a compound subject (2 or 3 topics) must be represented under several main headings. If a book has several topics it may be better to list it under a broader main heading even though the heading may not represent the contents precisely.
Cocaine	Alternate treatment of multi-topic works. Form headings or listing under as many headings as necessary.
	F. STANDARD LISTS AND THE LC LIST
Kiyoshikojin (Q4) Yamasaki	Catalogers' Aids: (1) Printed cards (2) subject heading lists. Use of standard list requires a knowledge of principles plus the ability to discriminate situations to which they apply.
Penpusher (Q2)	This is a professional rather than a clerical function, at least in the establishment of headings and syndetics.
Minerva (Q2)	Example.
Europa (Q4)	LC list is designed for LC. Other libraries should use it selectively.

Label and Q-type	Frame Topic Summary
Veroku (Q2)	Example: Makes of automobiles, and the Detroit public library.
	G. BUILDING A SYNDETIC NETWORK
Octavia	Restatement of the subject term selection process.
Achard (Q4)	Example: selection of "automobiles" or "motorcars" as a subject term.
Achellini	Access thru alternate terms "motor vehicles" and "motor cars" by 'see references.'
Abruzzi	Blind reference explained.
Ackerman (Q2)	'See refernces' not inconsistent with direct access.
Queenie.1 (Q2)	Syndetic concept. Linking of alternate terms for a topic to the main term by 'see references'.
Acton (Q4)	Books on the same topic share the same syndetic network but are required to be displayed (bibliographically listed) only under the main term.
Adair.1 (Q3)	This is Haykin's "unity". Problem: principal topic of a work not always easy to discern.
Sonya (Q4)	Problem: Books with misleading titles.
Ursula	Books with variant forms of subject terms in their titles, e.g., Muslims. Problem: What guides to use in assigning a heading? - Authority file and standard list.
Adelaide.1	Library environment factors which affect assignment of headings. Focus of course is on interpreting the L.C. structure.
	H. PROBLEMS OF ASSIGNING SUBJECT HEADINGS
Adelaide (Q10)	The number of see references under an unused alternate term. Moslems-Mohommedan mini-collection.
Adelon (Q10)	Selecting alternate terms for see references from titles in mini-collection.
NO LABEL	Cataloger not required to use all see references listed in LC in establishing a syndetic network.
Adler (Q2)	Problem: is frequency of term use in titles a basis for switching subject headings?

Label and Q-type	Frame Topic Summary
Agnissi (Q4)	Treatment of variant spelling (e.g. Muslims) not in LC list.
Africanus (Q4)	Methods of determining usage shifts and status of new terms.
Agnew (Q4)	Treatment of a locally used term (e.g. "Allahmites")
Crumpet (Q2)	Treatment of a new term not yet in LC e.g. lazars.
Adolphus.1	Summary of Moslem mini collection titles, main heading, and syndetic network.

SECTION TWO

A. SPECIFICITY IN SUBJECT HEADINGS

Albert	Specificity introduced.
Aconizio (Q7) Agriculture (Q7)	Formulating specific headings: India ink, Cluster-pines.
Agrippa (Q7) Ahlquist (Q7)	Formulating specific headings at a more inclusive level: Pines, Trees.
Ahrens (Q4)	Singular & plural term headings. Consistent use of each form necessary because of filing separation.
Akin (Q4)	Specificity reviewed.
Airy (Q4)	Level of specificity influenced by collection and user characteristics.
Akenside (Q1)	Applying specificity at a level greater than collection depth results in scattering, i.e., few books under each heading.

B. HOMONYMS AND FOREIGN PHRASES

Ainslie (Q7)	Homonyms qualified by parenthetical term, e.g., Lime (Tree)
Ackerblad (Q4)	Spelling and other factors quide main term selection. Overall guiding principle is user focus: i.e. the majority of users. Selection of English headings even for books on non-English terms, e.g., Aspen/ Espe/Tremble
Abasco (Q7)	Assimilated foreign phrases should be left untranslated.
Alcard (Q3)	Review of 'see reference' function.

C. REFERENCE FUNCTIONS AND SYMBOLS

Albani (Q3)	See also reference function described.
Subaru (Q3) Corona (Q4)	Practice in identifying see and see also references and *x* and *xx* symbols as used in the LC list.
Bluebird (Q7) Carol (Q1) Fairlady (Q3)	Practice in using these symbols. Example: 1: speed-writing. Example 2: Diplomacy
Gladstone (Q3)	Summary statement of meaning of *sa* and *xx* symbols.

Label and Q-type	Frame Topic Summary
Albermarle (Q3) Alberoni (Q3)	Practice on *x* and *xx* with Pinboys.
Morton (Q4)	See also references are often made to coordinate topics.
Cromwell (Q3) Albizzi (Q3) Albonzo (Q2)	Practice on *see*, *sa*, *x* and *xx*. Given the heading and reference function-supplying the symbol.
Doshisha (Q3)	Using syndetic symbol with an expanding heading structure.
Ritsumei (Q3) Ryukoku (Q3)	Practice in supplying both headings and symbols. Statistics example.
Meidai (Q3) Hosei (Q3)	Flags and Yacht Flags example.
Alcuin (Q3)	Example of *sa reference* between headings of coordinate scope.
Kent (Q4)	Sa reference not usually made from subordinate to superordinate topics.

D. ADJECTIVE AND NOUN HEADINGS

Juice (Q6)	Summary statement of user focus, direct access, and natural language. Introduction of adjectival heading in normal order.
Juicy.1	Use of see reference from inverted to uninverted adjectival headings.
Taupe (Q4)	Inverted forms are often not entered as alternate headings where noun is non-distinctive.
Gilroy (Q2)	Where the inverted or uninverted form sound equally natural, the form which places the more distinctive term to the fore is usually employed.
Percy (Q7)	Inversion is often used where the subject is vast and diversified, in order to reduce large blocks of titles all under the same entry term.
Danton (Q7)	Practice example: Military research.
Westby (Q7)	Simple adjective noun headings are often used rather than a more awkward prepositional phrase.
Angelus	2 nouns naming separate topics may be joined to name a 3rd topic.

Label and Q-type	Frame Topic Summary
Mentor (Q4)	Nouns in phrases may be used to refer to overlapping topics often treated together.
	E. SUBDIVISION IN GENERAL
Carlyle (Q3)	Restatement of specificity.
Alison (Q3)	An example of need for specificity.
Algardi	Aspects of a defined subject used as subheadings.
Allix (Q8)	Double dash entry style of subdivided headings.
Ainwick (Q2)	Subheading, subdivision and hierarchical classification.
Limit (Q4)	Subheadings as limitation of the subject in a way which produces more access points and smaller groups of records.
	F. FORM SUBDIVISION AND SUBHEADINGS
Spokane.1 (Q4)	Form subdivision defined and illustrated.
Ecoform	Student checks LC "Subdivisions of general application"
Manual (Q4)	Cataloged material as a basis of guidance in applying subdivisions. - handbooks, manuals, etc.
Elephant.1 (Q4)	Applying form subheadings on the basis of title.
Coller.1 (Q1)	Form subdivision is optional.
Altdorfer	Subdivision used with deep collections on noun topics e.g. Forgery of works of art.
	(Other types of subdivisions)
Lemon (Q4)	Brief introduction to other types of subdivision, e.g. chronological, geographical, topical. Problem 1: Trees--Diseases.
Hawthorne (Q4)	Problem 2: Molluscs--Dictionaries.
Buck (Q4)	Problem 3: Georgia--History--1775-1865.
Pottery (Q4)	Problem 4: Cathedrals--France.
Stone (Q4)	Problem 5: France--Administration.
	(Return to various forms of Subheads)

Label and Q-type	Frame Topic Summary
Yodoyabashi (Q7)	Examination of some commonly used form subheads. Heading Formulation 1: Chemical abstracts.
Brisket (Q7)	Heading formulation 2: Lectures on journalism.
Althorp (Q3)	Distinction between--"Bibliography" and--"Bio-bibliography."
Alvarez (Q7)	Heading formulation using "--Collection", given title and entry term.
Nakayamadera (Q2)	Description of material under "--Dictionaries"
Amadeo (Q7)	Heading formulation problem: "--Exhibitions".
Amato (Q2)	Use of "societies" explained.
Amerigo (Q4)	Use of "--Study and teaching" explained.
Ames	Summary of form subheads discussed.
Remember (Q10)	Student asked to supply name of previously mentioned types of subdivision.
G.	CHRONOLOGICAL SUBDIVISION
Verdi (Q3)	Chronological subdivision indicated by "history".
Amherst (Q2)	Factors governing use of dates in chronological subheadings.
Size (Q2)	Discussion of collection factor in use of dates.
Puccini (Q4)	Interpretation of chronological subheads.
Amici (Q7)	Topics with undefined origin or termination dates.
Ammon (Q4)	Precise dates not required in some subject areas.
Amundsen (Q7)	Omission of "--History" in subheads for important single events.
Bellini (Q4)	Mixed headings with and without "--History" in the same topic.
Andrada (Q4)	Historical events as main headings without dates, e.g., "Reformation." Headings for events with more than one name.
Tessai (Q10)	Access to the alternate name of the event.
Tosca (Q10)	Direct access in a chronologically subdivided heading.

Label and Q-type	Frame Topic Summary
	H. GEOGRAPHIC SUBDIVISION
Baader (Q7)	Geographic subdivision introduced.
Majuro (Q7)	Establishing geographic entities as independent headings.
Baan (Q2)	Comparison of two geographic headings formulated in above.
Babington (Q4)	Some headings not subdividable geographically because they are self-locating, too general, or too specialized.
Baez (Q3)	Problem 1: selection of geographically subdivided heading.
Bagnoli (Q4)	Problem 2: selection of geographically subdivided heading.
Bagshoow (Q1)	Meaning of "(direct)" instruction in LC list.
Bailey (Q5)	Meaning of "(indirect)" instruction in LC list. Exception for U.S. states.
Baird (Q7)	Problem: formulation of indirect heading for country plus local unit.
Bader (Q7)	Problem: formulation of heading for a small country without local unit.
Baer (Q7)	Direct headings most common in LC list.
Baker (Q7)	Local areas which are always used directly.
Toronto	Review (Optimal)
Mimi (Q7)	Rationale for areas always used directly.
Barkley (Q7)	Formulations of headings with country names.
Barlow (Q6)	Direct use of certain locales and direct headings in LC promote direct access.
Culture (Q2)	Access to direct forms when user looks under indirect forms first.
Honmachi (Q4)	Non governmental units which are always used directly.
Barnes (Q7)	Direct subdivision of local entities is often best regardless of LC directions for a heading.

Label and Q-type	Frame Topic Summary
Barrett (Q4)	Specifying country for local units which might be confused with different units of the same name.
	I. TOPICAL SUBDIVISION
Barry (Q10)	Introduction of topical subdivision.
Bartlett (Q6)	LC list as a source for topical subheads.
Amzot (Q4)	Situations which may require formulation of topical subheads not in LC or a standard list.
Bartolini (Q5)	Most topical subheads are useful in a variety of topic areas.
Kalif (Q2)	Subheads not in the LC list may be formulated but other lists should be checked as well.
Obayashi (Q4)	If a choice is available between a topical subhead and an unsubdivided adjective-noun form the latter is preferred.
Boasilicus (Q3)	Possible scattering of the subject may be a factor in selecting a form of the unsubdivided heading.
Basker (Q9)	In some cases inverting the order of the heading and topical subhead may be an equally plausible form.
Basil (Q6)	Problem of place vs. topic as entry term in geographically defined topics.
Bassano	Review option presented.
	J. PROPER NAMES
Bateman (Q3)	Proper names as subject headings.
Baudry (Q4)	Examples of the varieties of proper names.
Baumbach (Q5)	Categories of proper names.
Kwangaku (Q4)	The extended meaning of corporate bodies.
Baumgarten (Q3)	Most proper names are entered in direct natural language order. Western personal names are inverted.
Beale (Q6)	Non-western names are not always inverted, e.g. Chinese.
Bavius (Q3)	Dates as a distinguishing device.
Beach (Q2)	Personal names in titles are not changed in any way.

Label and Q-type	Frame Topic Summary
Beard (Q3)	Distinction between names as authors and names as subjects.
Beattie (Q3)	Proper names are almost entirely omitted from LC and other standard lists.
Beaufort (Q4)	Subdivision of proper name subject headings seldom done.
Beaver (Q3)	Exception: major figures with a large literature.
Bede (Q10)	Use of subdivision, even for major figures, is based on collection depth.
Bedford (Q3)	In LC LIST subdivision is displayed for only a few figures, to show typical patterns.
Beacher (Q5)	Most proper names are not in the LC LIST but are in the LC Catalog as subject and author headings. Some names in the LC LIST as parts of phrase headings.
	K. MAIN ENTRY AND SUBJECT ACCESS
Belasco (Q1)	Definition of main entry. Uniqueness of each to a work.
Bellybutton (Q6)	The heading in a main entry.
Bellingham (Q6)	Added entries.
Belloc (Q4)	Authors' names most commonly used as main entry terms.
Belmont (Q3)	Distinctiveness of author heading. Title and subject headings for added entries.
Beltrami (Q4)	Problem 1: Selection of main entry heading from three possible headings.
Brembo (Q2)	Works with identical title and subject.
Kandai (Q2)	Divided catalog and works with identical author and subject.
Benchley (Q5)	Problem 2: Selection of main and added entries.
	L. TRACINGS AND ADDED ENTRIES
Bandix (Q2)	Role of tracings in control of entries.
Kandinsky (Q10)	Differentiating subject from other added entries.
Tracing (Q4)	Problems in identifying various types of tracings. Problem 1: subject added entry.

Label and Q-type	Frame Topic Summary
Bengurion (Q4)	Problem 2: subject added entry.
Benoit (Q4)	Problem 3: subject added entry.
Benson (Q4)	Problem 4: joint author added entry.
Bentlink (Q4)	Problem 5: translator added entry.
Bentley (Q4)	Problem 6: title added entry.
Benton (Q4)	Problem 7: series added entry.
Berengaria (Q4)	Significance of absence of tracings.
Bergdorf (Q6)	Possibility of author added entry.
Beresford (Q10)	Subject headings for literary works.

SECTION THREE

Label and Q-type	Frame Topic Summary
	A. GENERAL PRINCIPLES AND THE SMALL LIBRARY
Southey (no question)	Turning to consideration of the small library.
Bernadette (Q4)	Why are LC methods useful for other libraries? Four different reasons offered for multiple choice.
Bernstein (Q2)	Reason #1 elaborated. (LC has encountered and acted upon almost every imaginable cataloging problem.)
Bilbo (Q2)	Reason #2 elaborated. (LC has a large staff.)
Berry (Q2)	Reason #3 is elaborated. (LC's response to change in usage.
Bertrand (Q2)	Need for adapting LC method to fit a particular collection.
Besant (Q4)	Same principles pervade.
Bessemer (Q4)	Introducing the Sears List.
Betterton (Q3)	(Reinforces *Besant*)
Harbin (Q4)	Reformulation of certain principles. Which ones are fully realizable? (None). Discussion of each if required.
Macao (no question)	Summary of *Harbin*.
Bewick (Q3)	Serviceability of Sears List for small libraries.
Biddle (Q9)	Which list might contain an explanation of rules?
Bienville (no question)	Sears list conveniences, e.g., blank space for annotation.
Matkin (Q2)(Q3)	Defining "authority file."
Bridewell (Q3)	Sears statement on specificity.
Bierstadt (Q4)	Useful limits of specificity. What is the criterion?
Billroth (Q7)	Problem: emperor penguins.

Label and Q-type	Frame Topic Summary
	B. METHODS OF ABRIDGEMENT USED BY SEARS
Bingham (Q2)	Terms not explicit in Sears may still be constructed according to Sears rules. Example: terns and gulls.
Juno (Q2)	*sa reference* cutoff.
Binne (Q4)	Related headings sometimes omitted in order to (3 choices.)
Birdwood (Q4)	Justified how? (7 choices.)
Yedo (Q4)	Abridgement by converting many headings to *see references*.
Blackett (Q2)	What is synonymy?
Blackstone (Q3)	Choice of synonym to use as subject term.
Blake (Q4)	Sears and current usage.
Snake (Q3)	Abridgement by omitting *see references*.
Blavatsky (Q4)	Synonymy vs. mutuality. Adapting Sears treatment to suit own situation.
Bleriot (Q7)	Problem: Make heading for a directory of clothiers.
Blondin (Q4)	Caution against blind reference.
Bliss (Q2)	More on synonymy.
Hepburn (Q9)	Recapitulation of Sears abridgement.
Boabdil (no question)	Other Sears instructions noted.
Bodley (Q3)	Sears coverage of adjectival phrase headings.
Bodmer (Q4)	Compound headings reviewed.
Glaubus (Q3)	Sears limited use of same. Choice of order of terms.
Boethius (Q10)	Problem: Identifying a compound heading.
	C. DRILL AND PRACTICE WITH ADJECTIVAL PHRASE HEADINGS, INVERSIONS, ETC.
Boldrewood (Q7)	Solar System
Bollingbroke (Q7)	Vocal music.

Label and Q-type	Frame Topic Summary
Bollivar (Q7)	Musical form.
Bonheur (Q7)	Child labor.
Boniface (Q7)	Unemployment insurance.
Todai (Q7)	Applied mechanics.
Bonvalle (Q7)	Life insurance.
Boone (Q7)	French sculptors; edible plants.
Bosch (Q7)	Dominion of the sea.
Tramp (Q3)	Professions explicit in LC, suggested in Sears.
Bosworth (Q4)	More on *Tramp*.
Bothwell (Q3)	Reminder to provide references.
Bottomley (Q1)	Need to have both lists at hand.
	D. ADVANCED SUBDIVISION
Retread (Q10)	Review of subdivision.
Boullion (Q2)	Evaluation of subdivision in terms of specificity, and in terms of direct access.
Borassa (Q2)	Evaluation in terms of scattering.
Klampus (Q3)	Example of *Borassa*.
Bourget (Q1)	Evaluation in terms of classificatory effect. Example.
Boutwell (Q1)	Evaluation in terms of self-evidence.
Zagreb (Q1)	Evaluation in terms of naturalness. Example.
Bovadilla (Q4)	Summary.
Bovary (Q3)	Instructional features of LC and Sears compared.
Bowditch (Q5)	More on *Bovary*.
Bowdoin (Q3)	Typical subdivision structures for nations, cities, founders of religions, etc.
Boyd (Q2)	Use of Sears for further examples of above.
Brabizon (Q3)	Geographic subdivision in Sears.

Label and Q-type	Frame Topic Summary
Bracton (Q3)	Headings unsuitable for geographic subdivision (LC).
Braddock (Q10)	Dispensing with geographic subdivision in small libraries.
Brahms (Q4)	Caution in above.

E. NATIONALITY AND ETHNIC QUALIFYING TERMS

Braithewaite (Q3)	Inversion as a means of avoiding scattering.
Brahmah (Q3)	Inversion not used for expatriates.
Brandeis (Q8)	Review of earlier statement regarding expatriates. Student example called for.
Branko (Q4)	Products of nationals.
Branmuffin (Q3)	Pluralizing: painting vs. paintings.
Brantome (Q7)	Practice: African artists.
Braque (Q7)	Practice: African art.
Brasidas (Q7)	Practice: Paintings in Switzerland.
Braxun (Q7)	Practice: English paintings.
Buffo (Q8)	Student example required.